Herd Behavior, Tail Risk Exposure and Asset Prices

Maxime Nicolas

Université Paris I Panthéon-Sorbonne
Ecole doctorale d'économie

Thèse en vue de l'obtention du titre de Docteur en Sciences Economiques

présentée et soutenue publiquement par:

Maxime Nicolas

le vendredi 20 janvier 2023

Herd Behavior, Tail Risk Exposure and Asset Prices

Thèse dirigée par:

Catherine Bruneau, Professeure à l'Université Paris 1 Panthéon-Sorbonne

JURY:

Christophe Boucher,	Professeur à l'Université Paris Nanterre	Rapporteur
Catherine Bruneau,	Professeure à l'Université Paris I	Directrice de Thèse
Gunther Capelle-Blancard,	Professeur à l'Université Paris I	Examinateur
Matthieu Garcin,	Enseignant chercheur à l'ESILV	Examinateur
Christophe Hurlin,	Professeur à l'Université d'Orléans	Rapporteur
François Longin,	Professeur à l'ESSEC	Examinateur

Contents

Résumé **3**

Abstract **4**

Remerciements **5**

Introduction (Français) **7**

Introduction (English) **14**

1 Estimating a model of herding behavior on social networks **20**

1.1 Introduction . . . 21
1.2 Methodology . . . 23
1.2.1 Agent-Based Model of Investor Interactions . . . 23
1.2.2 Network Structure . . . 24
1.2.3 Model Estimation . . . 25
1.3 Empirical Data . . . 27
1.3.1 Sentiment Index Estimation . . . 27
1.3.2 Stock Market Data . . . 30
1.4 Empirical Results . . . 33
1.4.1 Model Estimation . . . 33
1.4.2 Herding Intensity and Volatility Levels . . . 36
1.5 Discussion and Conclusion . . . 38
1.A Appendix . . . 39
1.A.1 Fokker-Planck Equation . . . 39
1.A.2 Finite Difference Method . . . 40

2 Nonparametric estimator of the TDC **43**

2.1 Introduction . . . 44
2.2 Tail-dependence coefficient . . . 45

2.3 Nonparametric estimation of the TDC . . . 46
2.3.1 Lower tail . . . 46
2.3.2 Upper tail . . . 48
2.3.3 Average of estimators . . . 49
2.4 Selection of the threshold . . . 51
2.4.1 Plateau-finding algorithm . . . 51
2.4.2 Minimization of the MSE . . . 52
2.4.3 Minimizing an average MSE . . . 57
2.5 A simulation study . . . 58
2.5.1 Gumbel simulations . . . 59
2.5.2 Student simulations . . . 60
2.5.3 Rotated Clayton simulations . . . 60
2.5.4 Gaussian simulations . . . 61
2.5.5 Analysis . . . 62
2.6 Empirical application . . . 62
2.7 Conclusion . . . 63
2.A Appendix . . . 64
2.A.1 Proof of Theorem 2.3.1 . . . 64
2.A.2 Proof of Proposition 2.3.1 . . . 65
2.A.3 Proof of Theorem 2.3.2 . . . 66
2.A.4 Proof of Proposition 2.3.2 . . . 66
2.A.5 Proof of Theorem 2.3.3 . . . 67
2.A.6 Uniform integrability condition of Theorem 2.3.1 for the independent copula 67
2.A.7 Tables of results for the simulation study . . . 70

3 Spurious tail risk factors and asset prices 74
3.1 Introduction . . . 75
3.2 The TDC and bias . . . 77
3.2.1 The TDC and the copula . . . 77
3.2.2 Estimation of the TDC . . . 79
3.2.3 Bias in the estimation of the TDC . . . 80
3.3 Data & variables . . . 82
3.3.1 Data . . . 83
3.3.2 Variables . . . 83
3.4 Empirical analysis . . . 84
3.4.1 Descriptive statistics and temporal variation . . . 85
3.4.2 Portfolio sorts . . . 87
3.4.3 Multivariate analysis . . . 87
3.4.4 Additional controls . . . 91
3.5 Conclusion . . . 92
3.1 Multivariate TDC in terms of copula functions . . . 94

3.2 Appendix simulation studies and estimation 95

Conclusion **96**

List of Tables

1.1 Social sample summary statistics. 30
1.2 Stock market data statistics. 31
1.3 Maximum likelihood estimation for M1: $U = \alpha_0 + \alpha_1 x$. 34
1.4 Maximum likelihood estimation for M2: $U = \alpha_0 + \alpha_1 x + \alpha_2 y$. 35
1.5 Maximum likelihood estimation for M3: $U = \alpha_0 + \alpha_1 x + \alpha_2 y \times z$. 36

2.1 Estimation methods. 59
2.2 Lower TDC: US vs other developed countries. 63
2.3 Upper TDC: US vs other developed countries. 64
2.4 Upper tail dependence with 100 Gumbel simulations. 70
2.5 Upper tail dependence with 100 Student simulations, with $\rho = 0$. 71
2.6 Upper tail dependence with 100 Student simulations, with $\rho = 0.25$. 71
2.7 Upper tail dependence with 100 rotated Clayton simulations. 72
2.8 Upper tail dependence with 100 Gaussian simulations. 73

3.1 Major tail risk exposure studies. 78
3.2 Summary statistics and correlations. 86
3.3 Equally weighted portfolio sorts based on correlation and tail risk. 88
3.4 Fama and MacBeth (1973) regressions. 89
3.5 Fama and MacBeth (1973) regressions and quintile portfolio on correlation. 91

List of Figures

1.1 Transient density function for various parameter values. 27
1.2 Numerical approximation of the transient density function. 27
1.3 Weekly financial returns and sentiment index. 32
1.4 Weekly financial returns and sentiment index. 32
1.5 Estimated parameters and the standard deviation of returns. 37
1.6 Dynamic estimation of M3 and the standard deviation of returns. 38
1.7 Finite difference scheme. 42

2.1 Correlation $\rho_{i,j}$ between $\widehat{\lambda}_{L,n}(i/n)$ and $\widehat{\lambda}_{L,n}(j/n)$. 51
2.2 Function of $\log\left(\widetilde{MSE}(i, C_{\psi(i/n)})\right)$. 55
2.3 Optimal threshold $\phi(\theta)$. 56
2.4 Threshold selection for the two-step plug-in approach. 57
2.5 RMSE for the upper tail dependence with 100 Gumbel simulations. 60
2.6 RMSE for the upper tail dependence with 100 Student simulations. 61
2.7 RMSE for the upper tail dependence with 100 rotated Clayton simulations. 61
2.8 RMSE for the upper tail dependence with 100 Gaussian simulations. 62

3.1 Gaussian copula versus Clayton copula. 80
3.2 Bias estimation in Gaussian simulations. 81
3.3 Sensitivity of the TDC by threshold selection and correlation. 82
3.4 Multivariate TDC bias as a function of the correlation coefficient. 82
3.5 Aggregated behavior of tail risk exposure and return correlation over time. 85

Résumé

Cette thèse de doctorat s'articule en trois chapitres qui traitent des comportements moutonniers, de la mesure de la sensibilité aux krachs financiers et de son impact sur les rendements financiers. Chacun de ces articles apportent une contribution méthodologique autour du risque de marché. Plus précisément, ils constituent des éléments de réponse à ces trois questions : Comment se forme le risque sur les marchés ? comment et avec quelle mesure peut-on mesurer le risque ? Est-ce que le risque est incorporé dans les rendements financiers ?

Le premier chapitre consiste en l'estimation d'un modèle d'agent décrivant la contagion du sentiment des investisseurs qui interagissent entre eux. A l'aide d'une analyse textuelle des messages publiés sur le réseau social StockTwits, la dynamique du sentiment des investisseurs est estimée pour des actifs financiers et des cryptos monnaies parmi les plus discutés sur la plateforme. Ce premier chapitre apporte des éléments d'observation empiriques sur le lien entre l'intensité de la contagion et le niveau de risque sur les marchés.

Le deuxième chapitre développe une méthode théorique pour l'estimation de la dépendance de valeurs extrêmes, en finance, cette mesure traduit la probabilité d'observer un krach simultané entre deux actifs. La crédibilité de la méthode est testée sur différentes données de simulations pour comparer sa performance par rapport à d'autres estimateurs. Une application sur la dépendance des valeurs extrêmes entre les rendements de l'indice des marchés US et les rendements des indices de marchés internationaux est également donnée.

Le troisième chapitre propose de réexaminer des études récentes qui mesurent l'effet de la sensibilité aux krachs sur les rendements des actifs financiers. La sensibilité aux krachs est donnée par la dépendance de valeurs extrêmes étudiée au chapitre précédent. Après avoir étudié un potentiel biais dans les estimations paramétriques et non paramétriques qui apparaissent lorsque qu'on a de fort niveau de corrélation, les précédentes études sur les rendements financiers sont reproduites. En incorporant le biais dans les analyses des précédents papiers, les effets significatifs sur les rendements précédemment documentés disparaissent.

Mots clés : Modèle d'Agent, Sentiment des Investisseurs, Comportements Moutonniers, Réseaux Sociaux, Coefficient de Dépendance de Valeurs Extrêmes, Estimations Non-Paramétriques, Copula, Analyse des Valeurs Extrêmes, Pricing d'actifs

Abstract

This thesis is constituted of three chapters on herding behaviors, the measure of crash risk sensitivity and its impact on financial asset returns. Each of this article provides a methodological and empirical contribution about risk in financial markets. More precisely, they provide answers to the following questions: Where does the risk in financial markets come from? How to measure the risk in financial markets? Is the risk incorporated in financial returns?

The first chapter is dedicated to the estimation of an Agent Based Model that describes herding behavior in the formation of investors sentiment. With the use of textual analysis of investors messages published on the social media StockTwits, the dynamic of investors sentiment is estimated for some of the most discussed financial asset and cryptocurrencies on the platform. This first chapter provides empirical observation that high levels of contagion are associated with higher levels of risk observed in financial returns.

The second chapter develops a theoretical method to estimate the extreme value dependency, in finance, this measures the probability of a joint crash between two financial assets. The usefulness and performance of the proposed estimator is tested in a simulation framework against multiple distributions. An empirical application is provided to measure the extremal dependence between the returns of the US market and other international markets returns.

In the third chapter we proposed to reexamine recent studies that provide evidence that the crash sensitivity impacts financial returns. The crash sensitivity in those corresponding studies is measured with the extremal dependence, presented in the previous chapter, between the asset returns and the market returns. After having presented potential biases in the parametric and non-parametric estimators when there is a high level of correlation, we replicate the studies dedicated to demonstrate the relationship between crash risk exposure and future excess returns. We proceed by showing that these results do not hold when we control for the correlation coefficient and other past returns behavior.

Keywords: Agent-Based Model, Investor Sentiment, Herding Behavior, Social Network, Tail dependence coefficient, Nonparametric estimation, Copula, Extreme values, Asset Pricing

Remerciements

Je souhaite exprimer mes remerciements et ma reconnaissance aux personnes sans qui ce projet n'aurait pu voir le jour.

Je tiens à remercier ma directrice de thèse qui m'a fait confiance et donné l'opportunité de réaliser ce travail de recherche. Je la remercie pour son exigence et son expertise qui ont contribué à la qualité de mes travaux. Je la remercie particulièrement pour la confiance et la liberté qu'elle m'a accordée dans la direction de ma recherche.

Je souhaite exprimer ma profonde reconnaissance aux Professeurs Christophe Hurlin et Christophe Boucher, qui ont accepté d'être les rapporteurs de cette thèse. Je les remercie pour leur travail de relecture, leurs commentaires et suggestions pertinentes lors de ma pré-soutenance de thèse qui ont grandement amélioré mes travaux. Je remercie également François Longin d'avoir accepté d'être membre de mon jury de thèse. C'est un immense honneur de soutenir devant un tel jury d'experts.

J'adresse un remerciement spécial à Matthieu Garcin. Je le remercie tout d'abord de m'avoir donné l'envie de faire de la recherche alors que j'étais encore son étudiant. Je le remercie de m'avoir ouvert à des thématiques quantitatives et divers domaines interdisciplinaires en économie et en finance qui sont généralement absents du parcours traditionnel. Je lui dois également la partie la plus aboutie de mon travail de thèse au travers notre collaboration. Je le remercie encore pour ses relectures, pour son soutien et sa confiance depuis le début de mon parcours.

Je tiens à remercier Pierre-Charles Pradier, pour son soutien moral indispensable, ses commentaires pertinents sur mes recherches, pour son travail de relecture sur l'ensemble de mes chapitres mais aussi pour ses conseils personnels et ses encouragements.

Je suis reconnaissant envers tous les membres de l'Ecole d'Economie de la Sorbonne que j'ai pu rencontrer. Je remercie en particulier Thomas Renault pour nos échanges, ses relectures et ses suggestions pertinentes, je le remercie également pour ses conseils et son aide à la préparation de ce projet de thèse. Je remercie également Jorgen Vitting Andersen pour son soutien et ses commentaires pertinents. Je remercie Béatrice Boulu Reshef pour nos collaborations et nos échanges.

Je remercie mes amis et collègues doctorants avec qui j'ai pu partager mon quotidien qui m'ont soutenu et écouté. Je pense en particulier à Adham pour son aide précieuse et nos nombreuses discussions, mais aussi Adrien, Aref, Corentin, Maximilien, Moussa, Mona, Pierre et tous

mes autres collègues.

Je remercie également tous mes amis et tous ceux que j'ai pu rencontrer, que je ne pourrai pas tous citer ici, qui ont permis de prêt ou de loin à l'aboutissement de cette thèse. Je remercie tout particulièrement Alexis pour ses relectures et son soutien moral depuis le début de ce projet. Je le remercie pour ses encouragements et nos réflexions sur le sens de la vie qui ont grandement influencé mes travaux.

Je remercie mes parents, ma grand-mère, ma sœur et mon frère, malgré leurs trop nombreux "t'en es où avec ta thèse?" ils n'ont jamais douté de moi. Je les remercie infiniment pour leur soutien.

Enfin, je ne peux clôturer ces remerciements sans avoir une pensée toute particulière pour Lynda, la personne avec qui je partage ma vie. Elle m'a toujours soutenu et encouragé depuis le début, je la remercie d'avoir rendu ma vie de doctorant meilleure. Je n'ai pas les mots pour exprimer ma profonde reconnaissance mais j'espère la rendre fière en lui dédiant cette thèse.

Introduction (Français)

" Une évaluation conventionnelle, fruit de la psychologie collective d'un grand nombre d'individus ignorants, est exposée à subir des variations violentes à la suite des revirements soudains que suscitent dans l'opinion certains facteurs dont l'influence sur le rendement escompté est en réalité assez petite. Les jugements manquent en effet des racines profondes qui leur permettraient de tenir. Dans les périodes anormales notamment, lorsque la croyance à la continuation indéfinie d l'état actuel des affaires est particulièrement peu plausible, même s'il n'y a pas de raison formelle de prévoir un changement déterminé, le marché se trouve exposé à des vagues d'optimisme et de pessimisme irraisonnées, mais après tout compréhensibles en l'absence d'une base solide de prévision rationnelle. "

Théorie générale de l'emploi, de l'intérêt et de la monnaie. John Maynard Keynes, 1936

La théorie de l'efficience informationnelle des marchés (Fama, 1965) est toujours au cœur de l'industrie financière moderne. Elle suppose que l'information est parfaitement incorporée dans les prix des actifs financiers. Suivant cette hypothèse, seules des informations économiques considérables seraient susceptibles de provoquer des variations importantes de la valorisation des actifs financiers. De plus, dans un marché où les prix reflètent parfaitement l'information disponible, l'apparition de bulles financières, qui se traduisent par une variation excessive par rapport la valeur fondamentale, reste limitée.

Il est aujourd'hui difficile de donner encore du crédit à cette théorie tant les exemples de bulles et de crises financières sont nombreux. Les crises financières des XXe et XXIe siècles comme le krach de 1929, la bulle internet et la crise des subprimes en sont les exemples les plus frappants. Par ailleurs, la mondialisation croissante et l'augmentation de l'interconnexion entre les marchés internationaux, rendent les crises et les bulles financières généralisées de plus en plus fréquentes (Sornette, 2017). Chacune de ces crises est un appel à revoir les modèles traditionnels pour expliquer ces phénomènes (Bouchaud, 2008).

La théorie d'efficience des marchés est intrinsèquement liée à l'hypothèse de marche aléatoire. Du point de vue statistique, cette hypothèse stipule que les rendements sont représentés par des processus qui ne comportent que des fluctuations bénignes, les observations de larges mou-

vements sont considérées comme “anecdotiques” et traitées comme des données aberrantes. Ces croyances peuvent s'avérer dramatiques et elles sont notamment désignées comme la conséquence de la crise des subprimes[1]. En effet, pour les modèles qui servaient à mesurer les risques des produits financiers structurés liés aux crédits subprimes, la probabilité que plusieurs emprunteurs fassent défaut de manière simultanée était nulle. Suivant leurs hypothèses, la possibilité d'une crise généralisée n'existait pas. Cependant depuis Mandelbrot (1963), l'observation suggère que les rendements des actifs représentent trop de mouvements extrêmes pour qu'ils soient en adéquation avec cette hypothèse. Il est préférable de considérer que les mouvements extrêmes sont constitutifs des dynamiques des cours d'actifs et non pas des aberrations statistiques.

D'une autre part, la théorie financière classique repose sur l'hypothèse de rationalité des agents, figurée par l'homo œconomicus qui représente les agents comme rationnels qui maximisent leur utilité en toutes circonstances. Dans cette optique, les investisseurs forment des anticipations rationnelles en mobilisant toute l'information disponible. Keynes (1936) a largement exprimé ses doutes sur le fait que les investisseurs se reposent uniquement sur les informations dont ils disposent pour former leurs choix d'investissement. Il le fait notamment au travers sa célèbre analogie du “concours de beauté” pour illustrer le fonctionnement des marchés boursiers. Il fait alors le parallèle avec un investisseur qui a tout intérêt à ne pas prendre en compte ses goûts personnels mais à choisir plutôt en fonction du consensus et de la majorité. Depuis, l'idée a largement été reprise dans la littérature qui montre que les investisseurs sont régis par la psychologie de masse qui les pousse à analyser l'information disponible non pas de manière objective mais par imitation (Scharfstein and Stein, 1990).

Depuis les années 1990, de nombreuses études en finance comportementale ont introduit la psychologie des investisseurs comme élément déterminant des instabilités des marchés financiers (Barberis and Thaler, 2003). Les comportements grégaires en constituent un des éléments principaux permettant d'expliquer la formation de crises et de bulles. Plus généralement, la finance comportementale a permis, à l'aide de modélisation, d'expliquer les anomalies de marché par différents biais comportementaux. On peut citer les études qui rapportent que les investisseurs sont sujet à des vagues d'optimisme ou de pessimisme, tirant les prix au-delà des fondamentaux économiques (Baker and Wurgler, 2007).

Le risque en finance est plus important que le laisse entendre la théorie classique, et nous ne devons pas reproduire les erreurs antérieures qui ont empêché l'identification de mesures de risque cohérentes (Colander et al., 2009). Dans une tentative de limiter les occurrences et les effets des instabilités et particulièrement des événements extrêmes sur les marchés financiers, les acteurs de l'industrie financière et les régulateurs doivent sans cesse se munir de nouveaux outils pour comprendre comment se forme le risque et comment le mesurer de manière adéquate. C'est dans cette perspective et pour répondre à ces enjeux qu'ont été développés les travaux de recherche qui constituent la présente thèse.

Le but de cette dissertation est de délivrer trois contributions originales permettant de contribuer de manière novatrice à l'analyse du risque de marché. Dans un premier temps, compren-

[1]Voir l'article “Recipe for Disaster: The Formula That Killed Wall Street”, `https://www.wired.com/2009/02/wp-quant/`

dre pourquoi des instabilités apparaissent nécessite d'appréhender la théorie économique d'un point de vue du comportement collectif. C'est pourquoi le premier chapitre est une contribution méthodologique et empirique à l'analyse des comportements grégaires en partant de données révélant des informations sur les interactions entre investisseurs. Dans le deuxième chapitre nous proposons une méthode adéquate pour mesurer le risque de dépendance de valeurs extrêmes. La contribution est essentiellement théorique mais nous proposons une application de la méthode pour mesurer les risques extrêmes partagés par les marchés américains-actions et les autres marchés actions développés. Le troisième chapitre montre que les mesures de dépendance de risques extrêmes peuvent être affectées de biais. Fort de l'acquis théorique obtenu dans le second chapitre, nous montrons que ces biais peuvent avoir une incidence déterminante sur la répercussion de la "sensibilité aux krachs" sur les rendements d'actifs. Avec la conséquence que ces biais peuvent radicalement changer les conclusions quant à l'impact de cette sensibilité sur les rendements.

Chapitre 1

Dans une tentative de comprendre la formation du risque financier, ce premier chapitre intitulé "Estimating a model of herding behavior on social networks" propose d'étudier et de mesurer les comportements grégaires. Ces stratégies d'imitation, qui visent à adopter le comportement du plus grand nombre sans prendre en compte ses informations personnelles, font l'objet de nombreuses études qui cherchent à expliquer l'origine des instabilités financières. On peut citer les travaux pionniers de Shiller et al. (1984) qui ont suggéré que les choix des investisseurs étaient issus d'effets de mode et de tendances communes plutôt que des motivations individuelles. Ces comportements peuvent être à l'origine de boucles de rétroaction positive pouvant entraîner un emballement et se résoudre en bulles spéculatives et krachs (Lux, 1995; Sornette, 2017). Ils sont également associés à "l'exubérance irrationnelle" décrite par Shiller (2015) qui fait référence à un profond excès d'optimisme collectif qui entraîne les prix au-delà de la valeur fondamentale.

Ce chapitre propose une approche originale pour étudier ces comportements grégaires. Nous faisons référence à un modèle théorique permettant d'expliquer la formation de comportements mimétiques, en vertu desquels chaque investisseur forme son propre sentiment à partir de l'observation du sentiment des autres investisseurs (Weidlich, 1971; Lux, 1995). Le modèle théorique est ensuite confronté à des observations empiriques relatives au sentiment des investisseurs. Pour mesurer le sentiment, nous utilisons des sources de données très récentes, provenant d'internet et plus précisément de l'observation des interactions permises par les réseaux sociaux. En effet, grâce à l'apparition de nouvelles plateformes dédiées aux marchés financiers, les réseaux sociaux spécialisés sont devenus des médias d'information qui rassemblent de grandes communautés d'investisseurs.

L'étude du sentiment des investisseurs, mesuré à partir des discussions d'investisseur sur internet, n'est pas récente dans la littérature (Antweiler and Frank, 2004; Das and Chen, 2007) mais la place qu'on lui attribue dans l'industrie financière s'est révélée de plus en plus importante au

cours du temps[2]. Non seulement l'utilisation des réseaux sociaux a explosé mais aussi les nouvelles méthodes d'intelligence artificielle et de big data permettent le traitement et l'analyse des grandes quantités des messages échangés sur ces réseaux. Par ailleurs, de récentes études ont montré que ces échanges reflétaient bien un sentiment collectif chez les investisseurs, au point de permettre la prédiction des rendements futurs des actifs. (Sprenger et al., 2014; Chen et al., 2014; Renault, 2017).

Au-delà du débat sur leurs utilité pour prédire les rendements financiers, ces données sont une aubaine pour les recherches comportementales. En effet, elles nous permettent de confronter les modèles théoriques qui nous décrivent la formation du sentiment des investisseurs avec des observations empiriques. Aujourd'hui encore trop peu d'études sont consacrées à la validation des modèles théoriques par des études empiriques (Cipriani and Guarino, 2014). Ce premier chapitre propose donc de combler ce manque en proposant une méthodologie pour estimer un modèle théorique sur le rôle du sentiment à partir d'observations relatives aux échanges de messages. Ce chapitre apporte également une contribution empirique en donnant une interprétation des paramètres estimés sur les données observées. Ces paramètres estimés nous donne une idée de l'intensité de la contagion dans la formation du sentiment sur différentes actions américaines et crypto-monnaies.

Nous sommes ainsi en mesure de valider un modèle d'agent à partir de données de sentiment portant sur des actifs financiers individuels. Par ailleurs, nos estimations fournissent des valeurs de paramètres cohérentes et compatibles avec les hypothèses de la contagion dans la formation du sentiment. On confirme ensuite les arguments d'études récentes qui montrent que de forts niveaux de volatilités sont engendrés par des comportements moutonniers (Froot et al., 1992; Blasco et al., 2012; Wang and Wang, 2018). De fait, nos résultats montrent que la mesure de contagion proposée est significativement plus élevée lorsque l'on observe de forts niveaux de volatilité sur les actifs financiers. Par ailleurs, les résultats ayant été obtenus de manière dynamique, ils nous ont permis de cibler la période de bulle subie récemment par le marché des crypto-monnaies et de montrer que l'intensité de la contagion est forte au moment de la bulle mais décroît rapidement ensuite.

L'évidence de ces comportements grégaires sur les marchés des crypto-monnaies entraîne une exposition à un risque supplémentaire pour les investisseurs. Par conséquent, ses résultats appellent à une prise en compte de la contagion par les régulateurs pour adopter des mesures afin de limiter les comportements grégaires sur ces marchés. Ces résultats sont également importants d'un point de vue théorique puisqu'ils permettent de valider des modèles existants dans la littérature par de nouvelles observations empiriques.

Chapitre 2

Même si les modèles d'agent étudiés au cours du premier chapitre nous aident à comprendre l'émergence des faits stylisés observés sur les marchés, ils ne fournissent pas d'outils pour mesurer le risque de manière adéquate. C'est dans cette perspective que ce deuxième chapitre

[2] Voirhttps://www.bloomberg.com/professional/blog/can-get-edge-trading-news-sentiment-data/

"Nonparametric estimator of the tail dependence coefficient: balancing bias and variance" a été développé avec Matthieu Garçin.

Mesurer le risque nécessite de s'éloigner du paradigme traditionnel "gaussien", pour lequel les risques extrêmes apparaissent comme des valeurs aberrantes et peu probables. Si les changements extrêmes dans les prix sont négligés dans les modèles traditionnels, la probabilité que de tels événements violents s'observent de manière simultanée pour plusieurs actifs l'est tout autant. Pourtant, les risques extrêmes sont rarement issus d'un comportement isolé et surviennent plutôt à la suite d'un comportement collectif. Dans cette perspective, il est décisif de se représenter la dépendance entre différentes variables aléatoires pour décrire ce comportement collectif. Dans le cas le plus facile, la dépendance entre deux actifs ou marchés financiers est décrite par le coefficient de corrélation de Pearson. Or la corrélation ne prend en compte que la dépendance linéaire et se révèle évidemment totalement inadaptée en gestion des risques extrêmes dont les dépendances sont fortement non-linéaires; d'ailleurs la dépendance se révèle plus forte dans les périodes d'instabilité (Longin and Solnik, 2001; Patton, 2004). Dans une optique de proposer une mesure adéquate de la dépendance « extrême », on s'intéresse dans ce chapitre au coefficient de dépendance de queue de distribution (CDQ). Ce coefficient mesure la probabilité qu'une variable aléatoire observe une valeur extrême sachant qu'une autre variable aléatoire l'observe également. Dit autrement, il nous permet de calculer la probabilité que deux variables aléatoires subissent un choc de manière simultanée. En finance, cette mesure est largement utilisée pour mesurer la probabilité d'observer un krach affectant simultanément deux marchés internationaux ou deux actifs financiers (Malevergne and Sornette, 2003; Poon et al., 2004; Caillault and Guégan, 2005). En particulier, son utilisation est, de fait, importante dans une optique de diversification de portefeuille ; à ce titre, on peut citer (De Luca and Zuccolotto, 2011; Wang and Wang, 2018) qui développent une méthodologie pour créer des portefeuilles robustes tels que leurs actifs ne krachent pas ensemble.

L'estimation du CDQ peut se faire de manière directe selon une approche paramétrique. Cependant l'efficacité de ce type d'approche dépend largement du choix de la fonction de vraisemblance qui est postulée. Pour gagner en flexibilité, praticiens et académiques privilégient des méthodes non paramétriques. Dans sa version non paramétrique introduite par Joe (1997), l'estimation du CDQ nécessite de choisir un seuil arbitraire au-dessus duquel la probabilité d'observer des valeurs extrêmes jointes doit être calculée. Les méthodes existantes ont recours à des heuristiques ou à l'observation graphique pour sélectionner ce seuil (Frahm et al., 2005a; Schmidt and Stadtmüller, 2006; Caillault and Guégan, 2005). Cependant, aucun cadre théorique n'a été développé jusqu'à présent, pour déterminer une règle de sélection du seuil, alors que cette sélection est déterminante pour la qualité de l'estimation.

La principale contribution de cet article consiste précisément dans la proposition d'un cadre théorique permettant de sélectionner ce seuil. Le seuil est simplement choisi de façon à minimiser l'erreur quadratique moyenne (EQM), calculée en fonction du biais et de la variance de l'estimateur. La performance de l'estimateur est ensuite évaluée sur la base de simulations et comparée à celle des estimateurs utilisés dans les approches traditionnelles. Les résultats montrent la cohérence du nouvel estimateur proposé, sans que celui-ci n'apparaisse plus performant

que les autres estimateurs. La méthode d'estimation est enfin appliquée pour évaluer le CDQ entre les rendements de l'indice du marché action US et les rendements des marchés actions de 17 pays développés.

Au-delà de l'apport méthodologique, ce chapitre apporte un éclairage et des résultats intéressants pour les gestionnaires d'actifs qui souhaitent diversifier les risques extrêmes des actifs de leurs portefeuilles. Cette méthodologie peut également intéresser le régulateur qui cherche à appréhender les sources simultanées de risque pour contrôler et éventuellement contrer le risque systémique. Il est aussi important de souligner que les développements de ce chapitre ont des applications potentielles au-delà de la finance car le CDQ est appliqué dans différents domaines comme l'hydrologie et le climat (Tawn, 1988; Poulin et al., 2007; Aghakouchak et al., 2010) mais aussi l'astronomie (Scherrer et al., 2009; Sato et al., 2011).

Chapitre 3

La théorie financière nous enseigne que tout risque supplémentaire doit être compensé par des rendements supérieurs (Campbell, 1996). Une grande partie de l'économie financière est dédiée à la recherche de facteurs de risque qui permettent d'expliquer les rendements financiers. La probabilité d'observer un événement extrême doit être également prise en compte comme un facteur de risque à part entière. Si les risques de krachs de marchés financiers sont depuis longtemps pris en compte dans l'étude de l'espérance des rendements des actifs financiers (Roy, 1952), la série de chocs financiers subie ces 30 dernières années a poussé les investisseurs à appréhender les risques extrêmes comme facteur à part entière dans l'évaluation de la prime de risque des actions financières. D'après des développements théoriques récents, il est établi que la prime de risque des actifs doit intégrer une composante représentant le risque de krach (Rietz, 1988; Barro, 2006; Bollerslev and Todorov, 2011).

A la suite de quoi, de nouvelles mesures de risques appelées "sensibilités aux krach" ont été proposées très récemment dans la littérature (Chabi-Yo et al., 2018). Ce type de mesure exprime l'intensité de l'exposition d'un actif financier au krach de marché ou à plusieurs facteurs de risque prédéfinis. Cette liste de facteurs a notamment été défini par Chabi-Yo et al. (2021) et correspond aux facteurs de risque du modèle Fama and French (1995, 2015) à cinq facteurs avec le facteur "momentum" (Carhart, 1997) et le facteur BAB "betting-against-beta" (Frazzini and Pedersen, 2014). Pour un actif donné, cette exposition aux krachs est évaluée statistiquement comme la dépendance de valeurs extrêmes entre les rendements de l'actif et le rendement du marché associé. Ces études proposent des évaluations empiriques de la capacité de l'exposition aux risques extrêmes à prédire en partie les rendements futurs. Le troisième chapitre, intitulé "Spurious tail risk factors and asset prices", remet en question les résultats obtenus dans la littérature récente sur la capacité prédictive des facteurs d'exposition aux krachs.

Les développements du chapitre 2 sur l'étude de la dépendance de queues de distributions ont montré les limites de son estimation, plus précisément l'existence d'un biais dans l'estimation du CDQ lorsqu'on observe un fort niveau de dépendance dans la totalité de la distribution. Le CDQ capture alors une dépendance en partie nourrie par la forte dépendance globale, observée

lorsque la corrélation linéaire est forte.

Nous proposons d'abord des explications théoriques au lien existant entre le CDQ et le coefficient de corrélation, puis, à l'aide de simulations, une évaluation de l'importance du biais dans le cas gaussien. Dans un deuxième temps, nous reproduisons les résultats des précédentes études prouvant que l'exposition au krach mesurée avec le CDQ est un facteur contribuant à la prime de risque. Nous montrons que ces résultats sont remis en cause pour des niveaux différents de corrélation lorsqu'on considère plusieurs sous portefeuilles avec différentes valeurs du coefficient de corrélation. Nos résultats nous conduisent à penser que l'effet prédictif de la sensibilité aux krachs provient de ce lien avec la dépendance générale au marché. Au-delà de la contribution méthodologique, ce troisième chapitre apporte des informations importantes pour les praticiens et analystes qui cherchent à identifier des facteurs prédictifs des rendements financiers, plus particulièrement liés aux risques extrêmes, la démonstration de leur validité restant à trouver.

Introduction (English)

"A conventional valuation which is established as the outcome of the mass psychology of a large number of ignorant individuals is liable to change violently as the result of a sudden fluctuation of opinion due to factors which do not really make much difference to the prospective yield; since there will be no strong roots of conviction to hold it steady. In abnormal times in particular, when the hypothesis of an indefinite continuance of the existing state of affairs is less plausible than usual even though there are no express grounds to anticipate a definite change, the market will be subject to waves of optimistic and pessimistic sentiment, which are unreasoning and yet in a sense legitimate where no solid basis exists for a reasonable calculation."

The General Theory of Employment, Interest and Money. John Maynard Keynes, 1936

Nowadays, the efficient market hypothesis (Fama, 1970) is still the cornerstone of modern finance. This theory argues that all available information is perfectly reflected in the price of financial assets. Following this hypothesis, only few important economic news releases would be able to cause extreme mouvement in financial asset valuations. Moreover, in a perfect market where prices perfectly reflect the available information, the appearance of financial bubbles, which result in an excessive variation compared to the fundamental value, remains limited.

Today, it is difficult to give credence to this theory, since there are so many examples of financial bubbles and crises. The financial crises of the 20th and 21st centuries, such as the crash of 1929, the internet bubble and the subprime crisis are the most striking examples. Moreover, with increasing globalization and the growing interconnection between international markets, widespread financial crises and bubbles are becoming more and more frequent (Sornette, 2017). Each of these crises is a call to revise the traditional models to explain these phenomena (Bouchaud, 2008).

The efficient market hypothesis is intrinsically linked to the random walk hypothesis. From a statistical point of view, this hypothesis states that returns are represented by random processes that contain only minor fluctuations. Observations of large movements are considered "anecdotal" and treated as outliers. These beliefs can be dramatic, they are notably referred to as the

consequence of the subprime crisis[3]. Indeed, for the models that were used to measure the risks of structured finance products linked to subprime loans, the probability that several borrowers would default simultaneously was zero. According to their assumptions, the possibility of a generalized crisis did not exist. However, as far as Mandelbrot (1963), the observation suggests that asset returns represent too many extreme movements to be consistent with this hypothesis. It is preferable to consider that extreme movements are constitutive of asset price dynamics rather than statistical aberrations.

On the other hand, classical financial theory is based on the hypothesis of the rationality of agents, represented by homo œconomicus, who states that agents are rational beings who maximize their utility in all circumstances. From this perspective, investors form rational expectations by using all available information. Keynes (1936) expressed his doubts about the fact that investors rely solely on the information available to make their investment choices. He did this in particular through his famous analogy of the "beauty contest" to illustrate the functioning of stock markets. In his analogy, he draws a parallel with an investor who has every interest to not take into account his personal tastes but rather choosing according to the consensus and the majority. Since then, the idea has widely inspired the literature, which shows that investors are governed by mass psychology, which pushes them to analyze the available information not in an objective manner but by imitation (Scharfstein and Stein, 1990).

Since the 1990s, numerous studies in behavioral finance have introduced investor psychology as a determinant of financial market instabilities (Barberis and Thaler, 2003). The analysis of herd behavior is one of the main elements that explain the formation of crises and bubbles. More generally, behavioral finance has made it possible, with the help of modeling, to explain market anomalies by various behavioral biases. One can cite studies that report that investors are subject to waves of optimism or pessimism that pull prices beyond economic fundamentals (Baker and Wurgler, 2007).

Risk in finance is more important than conventional theory suggests, and we must not repeat past mistakes that have prevented the identification of consistent risk measures (Colander et al., 2009). In an attempt to limit the occurrences and effects of instabilities and particularly extreme events in financial markets, the financial industry and regulators must continually acquire new tools to understand how risk is shaped and how to measure it adequately. It is in this perspective and in response to these challenges that the research work that constitutes this thesis was developed.

The purpose of this paper is to provide three original contributions to the analysis of market risk. First, to understand why instabilities occur, it is necessary to consider collective behavior into the economic theory. This is why the first chapter is a methodological and empirical contribution to the analysis of herd behavior based on data revealing information on the interactions between investors. In the second chapter we propose an adequate method to measure the risk of extreme value dependence. The contribution is essentially theoretical but we propose an application to measure extreme risks shared by the US equity markets and other developed equity markets. The third chapter shows that measures of extreme risk dependence can be affected by

[3]See “Recipe for Disaster: The Formula That Killed Wall Street”, `https://www.wired.com/2009/02/wp-quant/`

biases. Building on the theoretical background obtained in the second chapter, we show that these biases can have a decisive impact on the measure of "crash sensitivity" on asset returns. With the consequence that these biases can radically change the conclusions about the impact of this sensitivity on returns.

Chapter 1

In an attempt to understand the formation of financial risk, this first chapter entitled "Estimating a model of herding behavior on social networks" proposes to study and measure herding behavior. These imitation strategies, which aim to adopt the behavior of the crowd without taking into account their personal information, are the subject of numerous studies that seek to explain the origin of financial instability. We can cite the pioneering work of Shiller et al. (1984) who suggested that investors' choices were the result of fads and trends rather than individual motivations. These behaviors can be the source of positive feedback loops that can lead to a boom and initiate speculative bubbles and crashes (Lux, 1995; Sornette, 2017). They are also associated with the "irrational exuberance" described by Shiller (2015), which refers to a profound excess of collective optimism that drives prices beyond their fundamental value.

This chapter proposes an original approach to study these herd behaviors. We refer to a theoretical model to explain the formation of mimetic behavior, whereby each investor forms his own sentiment from the observation of the sentiment of other investors (Weidlich, 1971; Lux, 1995). The theoretical model is then tested against empirical observations of investor sentiment. To measure sentiment, we use very recent data sources, coming from the Internet and more precisely from the observation of interactions allowed by social networks. Indeed, thanks to the appearance of new platforms dedicated to financial markets, specialized social networks have become information media that gather large communities of investors.

The study of investor sentiment, as measured by investor discussions on the Internet, is not new in the literature (Antweiler and Frank, 2004; Das and Chen, 2007) but its role in the financial industry has become increasingly important over time[4]. Not only has the use of social networks exploded, but also new methods of artificial intelligence and big data allow the processing and analysis of the large quantities of messages exchanged on these networks. Moreover, recent studies have shown that these exchanges do indeed reflect a collective sentiment among investors, to the point of allowing the prediction of future asset returns. (Sprenger et al., 2014; Chen et al., 2014; Renault, 2017).

Beyond the debate on their usefulness in predicting financial returns, these types of data are a boon for behavioral research. Indeed, they allow us to confront theoretical models that describe the formation of investor sentiment with empirical observations. Today, there are still too few studies devoted to the validation of theoretical models by empirical studies (Cipriani and Guarino, 2014). This first chapter therefore proposes to fill this gap by proposing a methodology to estimate a theoretical model on the role of sentiment from observations of message exchanges. This chapter also makes an empirical contribution by providing an interpretation of the estimated

[4] see `https://www.bloomberg.com/professional/blog/can-get-edge-trading-news-sentiment-data/`

parameters on the observed data that gives us an idea of the intensity of contagion in the formation of sentiment on different U.S. stocks and crypto-currencies.

We are thus able to validate an agent model using sentiment data on individual financial assets. Furthermore, our estimates provide parameter values that are consistent with the hypotheses of contagion in sentiment formation. We then confirm the arguments of recent studies that show that high levels of volatilities are generated by herding behaviors (Froot et al., 1992; Blasco et al., 2012; Wang and Wang, 2018). Indeed, our results show that the proposed contagion measure is significantly higher when high levels of volatilities are observed on financial assets. Moreover, since the results were obtained dynamically, they allowed us to target the bubble period recently experienced by the crypto-currency market and show that the intensity of contagion is strong at the time of the bubble but decreases rapidly afterwards.

The evidence of such herding behavior in crypto-currency markets results in additional risk exposure for investors. Therefore, its results call for regulators to take contagion into account in order to adopt measures to limit herd behavior in these markets. These results are also important from a theoretical point of view since they allow us to validate existing models in the literature with new empirical observations.

Chapter 2

Even if the agent models studied in the first chapter help us to understand the emergence of the stylized facts observed on the markets, they do not provide tools to measure risk adequately. It is in this perspective that this second chapter "Nonparametric estimator of the tail dependence coefficient: balancing bias and variance" has been developed with Matthieu Garçin.

Measuring risk requires going beyond the traditional "Gaussian" paradigm, for which extreme risks appear as unlikely outliers. If extreme price changes are neglected in traditional models, so is the probability of such violent events occurring simultaneously for several assets. However, extreme risks are rarely the result of isolated behavior, but rather occur as a result of collective behavior. From this perspective, it is crucial to represent the dependence between different random variables to describe this collective behavior. In the easiest case, the dependence between two financial assets or markets is described by the Pearson correlation coefficient. However, correlation only takes into account linear dependence, and is obviously totally unsuitable for the management of extreme risks, where dependence is highly non-linear; moreover, dependence is stronger in periods of instability (Longin and Solnik, 2001; Patton, 2004).

In an attempt to propose an appropriate measure of "extreme" dependence, this chapter focuses on the tail dependence coefficient (TDC). This coefficient measures the probability that a random variable observes an extreme value knowing that another random variable also observes it. In other words, it allows us to calculate the probability that two random variables will experience a shock simultaneously. In finance, this measure is widely used to measure the probability of observing a crash simultaneously affecting two international markets or two financial assets (Malevergne and Sornette, 2003; Poon et al., 2004; Caillault and Guégan, 2005). In particular, its use is, in fact, important from a portfolio diversification perspective; as such, we can cite (De Luca

and Zuccolotto, 2011; Wang and Wang, 2018) who develop a methodology to create robust portfolios such that their assets do not crash together.

The estimation of the TDC can be done directly using a parametric approach. However, the efficiency of this type of approach depends largely on the choice of the postulated likelihood function. To gain flexibility, practitioners and academics prefer non-parametric methods. In its non-parametric version introduced by Joe (1997), the estimation of the TDC requires the choice of an arbitrary threshold above which the probability of observing joint extreme values must be calculated. Existing methods use heuristics or graphical observation to select this threshold (Frahm et al., 2005a; Schmidt and Stadtmüller, 2006; Caillault and Guégan, 2005). However, no theoretical framework has been developed so far to determine a threshold selection rule, even though this selection is crucial for the quality of the estimation.

The main contribution of this paper consists precisely in the proposal of a theoretical framework for selecting this threshold. The threshold is simply chosen to minimize the mean square error (MSE), calculated from the bias and variance of the estimator. The performance of the estimator is then evaluated on the basis of simulations and compared to that of the estimators used in traditional approaches. The results show the consistency of the proposed new estimator, without a clear outperformance compared to the other estimators. The estimation method is finally applied to evaluate the TDC between the returns of the US equity market index and the returns of the equity markets indexes of 17 developed countries.

Beyond the methodological contribution, this chapter provides interesting insights and results for asset managers who wish to diversify the extreme risks of the assets in their portfolios. This methodology may also be of interest to the regulator who seeks to understand the simultaneous sources of risk in order to control and eventually counter systemic risk. It is also important to note that the developments in this chapter have potential applications beyond finance as the TDC is applied in different fields such as hydrology and climate (Tawn, 1988; Poulin et al., 2007; Aghakouchak et al., 2010) but also in astronomy (Scherrer et al., 2009; Sato et al., 2011).

Chapter 3

Financial theory teaches us that any additional risk must be compensated by higher returns (Campbell, 1996). Much of financial economics is devoted to finding risk factors that explain financial returns. The probability of observing an extreme event must also be taken into account as a specific risk factor. While the risks of financial market crashes have long been taken into account in the study of the expectation of returns on financial assets (Roy, 1952), the series of financial shocks experienced over the last 30 years has led investors to consider extreme risks as a factor of its own in the evaluation of the risk premium on financial stocks. According to recent theoretical developments, it is established that the risk premium of assets must incorporate a component representing crash risk (Rietz, 1988; Barro, 2006; Bollerslev and Todorov, 2011).

In this context, new risk measures called "crash sensitivities" have been proposed very recently in the literature (Chabi-Yo et al., 2018). This type of measure expresses the intensity of a financial asset's exposure to a market crash or to several predefined risk factors. This list of fac-

tors was notably defined by (Chabi-Yo et al., 2021) and corresponds to the risk factors of the Fama and French (2015) five-factor model with the momentum factor Carhart (1997) and the betting-against-beta factor (Frazzini and Pedersen, 2014). For a given asset, the crash risk exposure is statistically assessed as the extreme value dependence between the asset's returns and the associated market return. These respective studies offer empirical assessments of the ability of extreme risk exposure to partially predict future returns. The third chapter, entitled "Spurious tail risk factors and asset prices", questions the results obtained in this recent literature on the predictive capacity of the tail risk exposure factors.

The expertise acquired in Chapter 2 on the study of tail dependence has allowed us to identify the limitations of its estimation, more precisely the existence of a bias in the estimation of the TDC when a strong level of dependence is observed in the whole distribution. The TDC then captures a dependence partly fed by the strong global dependence, observed when the linear correlation is strong.

We first propose theoretical explanations for the link between the TDC and the correlation coefficient, and then, using simulations, we give an evaluation of the importance of the bias in the Gaussian case. In a second step, we reproduce the results of previous studies proving that the exposure to the crash measured with the TDC is a factor contributing to a risk premium. We show that these results are challenged for different levels of correlation when considering several sub-portfolios with different values of the correlation coefficient. Our results lead us to believe that the predictive effect of the crash sensitivity comes from this link to the overall market dependence.

Beyond the methodological contribution, this third chapter provides important information for practitioners and analysts who seek to identify predictive factors of financial returns, more particularly related to extreme risks, the demonstration of their validity remaining to be found.

1

Estimating a model of herding behavior on social networks

Published, *Physica A: Statistical Mechanics and its Applications* (2022)

This work has been presented at the following conferences:

- ▷ *28th Annual Conference of the Multinational Finance Society (MFS)*, Gdansk University of Technology, Poland
- ▷ *World Conference of the Economic Science Association (ESA)*, Massachusetts Institute of Technology, Boston/Cambridge MA, USA
- ▷ *28th International Conference Computing in Economics and Finance (CEF)*, Southern Methodist University, Dallas TX, USA

Abstract

In this paper, we estimate an agent-based model (ABM) to investigate herding behaviors in the formation of investor sentiment. We formalize a simple opinion dynamics model in a social network framework and rely on a numerical method to estimate its parameters. We derive a sentiment proxy from the weekly aggregation of online messages concerning 15 US stocks and 5 cryptocurrencies. Our empirical results suggest a strong impact of herding behavior on the formation of sentiment toward highly volatile assets. For such assets, we simultaneously find limited impacts of financial returns and investor attention on the opinion formation process, suggesting that investor sentiment is explained by social interactions. On the other hand, we find a limited influence of social interactions on sentiment regarding less volatile assets, whose formation process is instead explained by the strong influence of financial returns and investor attention. In particular, we find that herding behavior was significantly higher and played a major role in the sentiment formation process regarding cryptocurrencies when the bubble occurred.

Keywords: Agent-Based Model, Investor Sentiment, Herding Behavior, Social Network
JEL Classification: C13, C63, G12

1.1 Introduction

Herding behavior in financial markets was initially explained by Keynes' analogy of the beauty contest, described as traders' attempting to forecast "what average opinion expects the average opinion to be" and was later clarified as the action of "[conforming] with the behavior of the majority or the average" (Keynes, 1936, 1937). In light of these observations, Shiller et al. (1984) documented that investors spend a large proportion of their time reading about others' investments and gossiping about the successes or failures of their investments. This situation suggests that opinion formation is a social process governed by individual suggestibility and group pressure. With the current widespread access to social media, such behavior could be stronger than ever. In fact, online stock message boards have become a popular way for investors to inform themselves, discuss breaking news and corporate events, and comment on asset returns.

Herding behavior has been well documented in theoretical frameworks of agent-based models (ABMs). Most of these studies attempt to explain the known "stylized facts" observed in financial time series, such as clustered volatility and the so-called fat-tail property of asset returns (Lux and Marchesi, 1999, 2000; Cont and Bouchaud, 2000; Iori, 2002; Zheng et al., 2004), while some notable models focus more narrowly on investor opinion dynamics (Topol, 1991; Kirman, 1991; Banerjee, 1992; Orléan, 1995; Lux, 1995, 1998). Although these models can reproduce the empirical behavior of asset prices, they are mostly based on prior economic assumptions. Accordingly, empirical measurements would be welcome to strengthen theoretical analyses of herding behavior (Cipriani and Guarino, 2014). To this end, many approaches to validating ABMs with empirical data have recently been proposed[1]. Accordingly, a growing subset of the literature is focusing on the estimation of ABMs. The main concern of studies of this kind is to ensure that such models can accurately reflect real-world data.

This paper attempts to fill this gap by estimating a theoretical model of herding behavior based on empirical data. Our main contribution is to formalize the simple Weidlich-Lux opinion dynamics model (Weidlich, 1971; Lux, 1995) in a social network. Indeed, compared to previous authors Lux (2009, 2012); Shi et al. (2019) we formalize the model in a networked framework and obtain more consistent results. Moreover, previous studies presented estimates of herding intensity based either on an economic climate survey Lux (2009, 2012) or on stock markets indexes (Shi et al., 2019). However, there is still no evidence of herding in the sentiment index at the asset level. To address this issue, we use sentiment analysis and text mining techniques to derive a sentiment proxy from the weekly aggregation of online messages concerning 15 US stocks and 5 cryptocurrencies. We rely on numerical methods to estimate the parameters of a model of opinion formation. In line with previous authors who have found that volatility is driven by herding behavior (Froot et al., 1992; Blasco et al., 2012; Wang and Wang, 2018), we attempt to link herding intensity with the level of volatility. To the best of our knowledge, this is the first attempt at estimating an ABM based with a networked model. Our study is also related to recent research that tries to make sense of the boom and bust cycles in the cryptocurrency market by studying

[1] For discussions on the empirical validation of ABMs in economics, see Fagiolo et al. (2007) and Lux and Zwinkels (2018).

herding behavior (Bouri et al., 2019; da Gama Silva et al., 2019). However, our study is, to the best of our knowledge, the first to measure herding behavior in the formation of investor sentiment regarding cryptocurrencies.

Antweiler and Frank (2004) and Das and Chen (2007) were among the first authors to derive a proxy for investor sentiment from user-generated content in an attempt to explain stock returns. They used Yahoo! message boards and found no significant relationship with stock returns, volume, or volatility. Their results were later confirmed by Kim and Kim (2014), who showed that market sentiment was shaped by previous stock performance. In contrast, later studies showed that social media is capable of reflecting collective investor sentiment trends and has predictive power for future asset prices (Sprenger et al., 2014; Chen et al., 2014; Renault, 2017; Guo et al., 2017). Nonetheless, most research has been concerned with simply deriving a proxy to predict stock returns, whereas there is little empirical research that describes how investor sentiment itself is formed and evolves over time. A preliminary approach to this topic can be found in the estimation of ABMs of herding behavior in the formation of investor sentiment. Such models attempt to provide empirical validation of ABMs and a measure of herding intensity. Alfarano et al. (2005, 2008) estimated the parameters of a simple stochastic model of information transmission initially designed by Kirman (1993) to explain herding behavior in ant colonies. They first derived the underlying parameters of the model using a parametric approach and then linked the fat-tail property of the distribution of returns to the intensity of changes in strategy among traders. A relatively new branch of the literature focuses on the estimation of an interaction model, as initially proposed in the field of quantitative sociology by Weidlich (1971), to study the structure of social groups of individuals mutually influencing each other with respect to their decision behavior (Franke, 2008; Lux, 2009, 2012; Ghonghadze and Lux, 2011; Lux, 2018; Shi et al., 2019). Franke (2008) and Lux (2009) used survey expectations regarding economic growth[2] to estimate the parameters of models of social opinion formation among agents. Ghonghadze and Lux (2011) used EU business and consumer survey data for 12 European countries and assessed the fitness of a model with respect to its out-of-sample forecasting performance. Lux (2012) used weekly records of investor sentiment for the German stock market. More recently, Shi et al. (2019) used data on investor interactions on Chinese online forums to estimate contagion phenomena in sentiment formation for different industry sectors.

Notably, however, the main drawback of the Weidlich-Lux model is that it relies on two assumptions: that the number of agents N is relatively small and that the agents are embedded in a fully connected network, in which each agent interacts with every other agent. Alfarano and Milaković (2009) documented the lack of robustness of such models. Specifically, their results are not robust with respect to an increase in the number of agents. As the number of agents increases, a model of this kind becomes ill-defined; consequently, such models obviously cannot satisfactorily capture the interactions among the large numbers of participants found in real financial markets. Usually, these models consider every individual as the origin of the contagion process and pay little attention to the underlying network structure that defines the possibility of agent interaction. This finding is in line with recent studies that have documented the estimation of

[2]The ZEW Business Climate Index for the German economy.

a small "effective" number of agents (Lux, 2009). In the present paper, we attempt to overcome this issue of N-dependence documented by Alfarano and Milaković (2009) by accounting for the estimated underlying network structure in the model.

The main results of this paper can be summarized as follows:

- We formalize a simple ABM of opinion formation in social networks to overcome the problem of N-dependence.
- We estimate the herding parameters in the processes of sentiment formation with respect to 20 financial assets.
- We link the intensity of herding behavior with the volatility of the corresponding asset.

The paper is organized as follows. In Section 2, we formalize a simple ABM of sentiment formation in a network-based framework and develop a corresponding methodology for estimating the model parameters. Section 3 describes the data used to derive a sentiment index as well as the associated stock market data. Section 4 documents the estimation results. Section 5 concludes the paper.

1.2 Methodology

1.2.1 Agent-Based Model of Investor Interactions

Following Lux (1995, 1998), we adapt the model of opinion formation first introduced in the field of quantitative sociology by Weidlich (1971). In this general framework, it is assumed that only two opinions exist in the modeled society. Accordingly, in the following application scenario, agents are assumed to be investors who can be classified as having either an optimistic (bullish) opinion when they expect the price to increase or a pessimistic (bearish) opinion otherwise. The configuration of the population of investors $\{n_+, n_-\}$ at time t thus consists of two subgroups with occupation numbers n_+ and n_-, corresponding to optimistic (+) and pessimistic (-) opinions, respectively. The overall population size is $2N = n_+ + n_-$, and the average opinion of the investors is given by the following sentiment index:

$$x = \frac{n_+ - n_-}{2N} \quad \text{with } x \in [-1, 1] \tag{1.1}$$

Accordingly, $x > 0$ ($x < 0$) corresponds to a situation in which optimistic (pessimistic) investors are predominant, while $x = 0$ corresponds to a balanced situation. Agents may change their opinions over time and, in so doing, switch between the two possible subgroups (+) and (-). These switches occur based on individual transition probabilities that govern the overall dynamics of opinion formation. We assume a homogeneous population in which each individual has the same individual transition probabilities per unit time period, denoted by p_+ for the transition from (-) to (+) and p_- for the transition from (+) to (-). We define $n = (n_+ - n_-)/2$ to characterize

the configuration of the system. It is assumed that the transition rates take exponential forms:

$$\begin{cases} p_+(n) = v\exp(U) \\ p_-(n) = v\exp(-U) \end{cases} \tag{1.2}$$

where v is a time scale parameter that determines the frequency of switches between groups and $U(\cdot)$ is an influence function encompassing the factors that influence the rates of change of the sentiment transitions. In the basic Weidlich model, $U(\cdot)$ depends solely on the current population configuration, as described by the opinion index x:

$$U = \alpha_0 + \kappa n = \alpha_0 + \alpha_1 x \tag{1.3}$$

with $\alpha_1 = \kappa N$ and $Nx = n$. The two parameters of this model can be described as follows. (i) The constant bias factor α_0 reflects individual preferences toward an opinion, independent of the opinions of other people. The probability that an agent will change from opinion (-) to opinion (+) is increased when $\alpha_0 > 0$, correspondingly reducing the probability of changing from (+) to (-), whereas the opposite is true for a negative α_0. (ii) The contagion parameter α_1 (representing the herding effect) measures the intensity of sentiment contagion or herding behavior. It reflects group pressure influencing an individual in favor of the opinion of the majority. For a large positive α_1, the probability of transition in the direction of the majority opinion increases, and this effect increases with increasing $|x|$.

1.2.2 Network Structure

The basic Weidlich model assumes that individuals observe the current societal configuration at all times t when forming their opinions. Thus, it implicitly assumes the equivalent of a fully connected network, in which each agent can interact with every other agent, and it assumes that the contagion process is driven by the whole system configuration. Accordingly, the impact of sentiment on the individual transition probabilities expressed in (1.2) can be described by κNx. However, although this assumption may be true for small N, individuals can participate in only a limited number of interactions, and thus, as N increases, this assumption becomes less likely to hold. Assuming now that agents change their opinions under the influence of their neighbors, let us define the socioconfiguration of the neighbors of agent i as follows:

$$n(i,\mathbf{J}) = \left(n_+(i,\mathbf{J}) - n_-(i,\mathbf{J})\right)/2 \tag{1.4}$$

where $n_+(i,\mathbf{J})$ and $n_-(i,\mathbf{J})$ denote the numbers of i's neighbors that are in the optimistic (+) and pessimistic (-) states, respectively[3]. $\mathbf{J}$ represents the information about the network configuration between agent i and another agent j ($j \neq i$). To simplify the model, we employ a mean-field approximation as described by Alfarano and Milaković (2009) and assume that any heterogeneity in the neighbor configuration $\mathbf{J}$ is due solely to negligible fluctuations. Accordingly, we drop the

[3] $n_X(i,\mathbf{J}) = \sum_{j\neq i} D_X(i,j)$ is the number of neighbors of agent i in state X as defined in terms of an indicator function $D_X(i,j)$, which is equal to 1 if agent i is connected to agent j in state X.

term **J** and use the following approximation:

$$\langle n(i)\rangle = \left(\langle n_+(i)\rangle - \langle n_-(i)\rangle\right)/2 \tag{1.5}$$

Additionally, we further assume homogeneity in the neighbor configuration and replace the number of neighbors of each agent i with the average number of neighbors, D. In graph theory, this quantity is called the average degree of the nodes in the network. The degree reflects the likelihood that a node will receive information flowing through the network. The socioconfiguration of the neighbors of agent i can thus be expressed in terms of the average node degree D as follows:

$$\begin{cases} \langle n_+(i)\rangle = D\frac{n_+}{2N} \\ \langle n_-(i)\rangle = D\frac{n_-}{2N} \end{cases} \tag{1.6}$$

where $n_+/2N$ and $n_-/2N$ are used to approximate the "unconditional" probability that a neighbor of agent i is in state (+) or (-), respectively. For convenience, we adopt the notation $d = D/2$ and obtain the following approximation of the socioconfiguration of agent i's neighbors:

$$\langle n(i)\rangle = d\frac{n_+ - n_-}{2N} = dx. \tag{1.7}$$

At this microscopic level, the function U in the expressions for the transition rates can be written as

$$\langle U_i\rangle = \alpha_0 + \kappa\langle n(i)\rangle = \alpha_0 + \alpha_1 x \tag{1.8}$$

with $\alpha_1 = \kappa d$. The main difference between equations (1.3) and (1.8) lies in the effect of the herding parameter κ, which vanishes for $N \to \infty$ in (1.3). We now give a better insight to the effect of α_1, which now incorporates the coefficient D, to account only for a local interaction among individuals. The first contribution of this paper lies in this formalization, as it permits a more realistic model representation and enables us to estimate a suitable model for large N. While Alfarano et al. (2008) and Alfarano and Milaković (2009) have pointed out the analytical implications of different network topologies in the ant model of Kirman (1993), we consider a given network topology in which the average degree D is observed. In this paper, D is measured as the average number of potential interactions among traders in the considered microblogging platform.

1.2.3 Model Estimation

The probability distribution of the sentiment index is implied by the distribution of the ordered pairs $\{n_+, n_-\}$. Let us use $P(\{n_+, n_-\}, t)$ to denote the corresponding distribution function at time t, which we can further abbreviate as $P(x,t)$ for simplicity of expression. Since we assume that agents can continuously change their beliefs over time, the dynamics of the opinion index can be approximated in terms of continuous variables. Weidlich (1971) showed that the equation of motion can be expressed in the standard form of a one-dimensional Fokker-Planck equation (given in appendix A). Accordingly, $P(x,t)$ can be expressed in terms of the socioconfiguration of

agent i's neighbors $\{\langle n_+(i)\rangle, \langle n_-(i)\rangle\}$, as it is assumed that $p(\langle n(i)\rangle, t)$ follows the same dynamics

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}[A(x;\boldsymbol{\theta})P(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[D(x;\boldsymbol{\theta})P(x,t)] \tag{1.9}$$

with a drift coefficient

$$A(x;\boldsymbol{\theta}) = (1-x)p_+(dx) - (1+x)p_-(dx) \tag{1.10}$$

and a "fluctuation coefficient"

$$D(x;\boldsymbol{\theta}) = \frac{1}{d}\left[(1-x)p_+(dx) + (1+x)p_-(dx)\right] \tag{1.11}$$

The task is to estimate the parameters $\boldsymbol{\theta} = \{\alpha_0, \alpha_1\}$ from a sample of T observations $X_0, \cdots, X_T$. To obtain the parameter estimates, we first need to estimate the transient density function $P(x,t)$ of the sentiment index x at time t. Following Lux (2009, 2012), we apply a numerical maximum likelihood approach based on numerical solutions of the Fokker-Planck equation, as previously suggested by Jensen and Poulsen (2002) and Hurn et al. (2007). For estimation, we use a finite difference scheme based on the Crank-Nicolson method, as described in appendix B. Thus, the estimate of $\boldsymbol{\theta}$ is obtained by maximizing the log-likelihood function of the observed sample:

$$\log \mathscr{L}(\boldsymbol{\theta}) = P_0(x_0;\boldsymbol{\theta}) + \sum_{i=1}^{T} \log P(x_i \mid x_{i-1};\boldsymbol{\theta}) \tag{1.12}$$

where $P_0(x_0;\boldsymbol{\theta})$ is the initial density, which can be omitted when estimating the log-likelihood for the whole sample since it has only a small impact on the entire function, and $P(x_i \mid x_{i-1};\boldsymbol{\theta})$ is the value of the transitional probability density function (PDF) at (x_i, t_i) for a process starting at (x_{i-1}, t_{i-1}) and evolving at (x_i, t_i).

Figure 1.1 shows multiple equilibrium shapes of the transient density function for various values of the model parameters for $d = 25$, $x_{i-1} = 0$, and $v = 1.5$ with an increasing herding parameter (α_1); when this parameter is zero, the sentiment is concentrated around the α_0 parameter, and the dispersion increases with increasing α_1. The third graph in Figure 1.1 shows the typical "bimodal" density observed when the bias is null and the herding parameter is high. The displayed configuration illustrates that a balanced opinion configuration is improbable. Figure 1.2 shows two examples of numerical estimates of the transient density function with parameter values of $v = 1.5$ and $d = 25$; for the first graph, $\alpha_0 = 0.1$ and $\alpha_1 = 1.2$, while for the second graph, $\alpha_0 = 0$ and $\alpha_1 = 1.5$.

Figure 1.1: Transient density function for various parameter values.

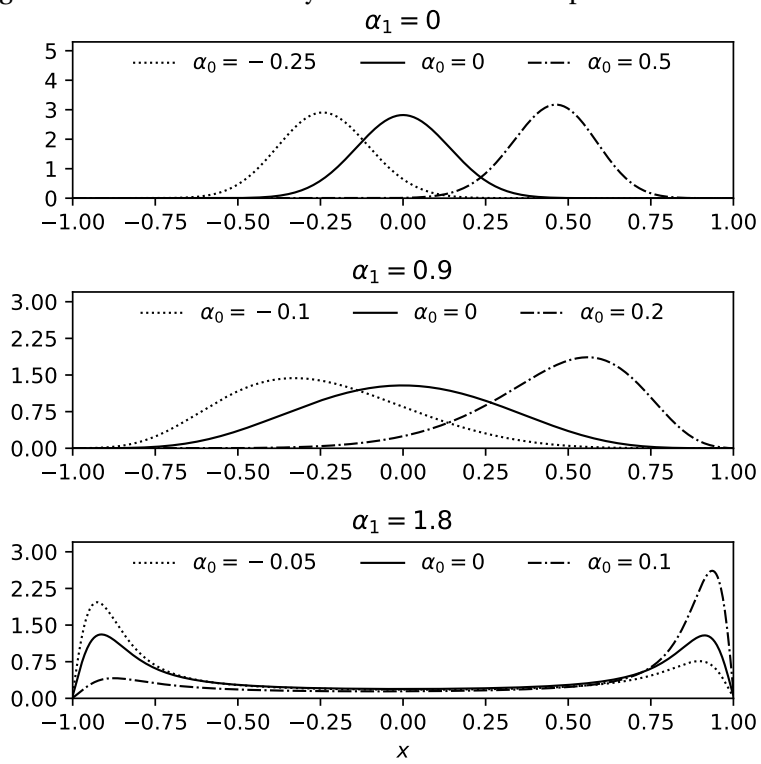
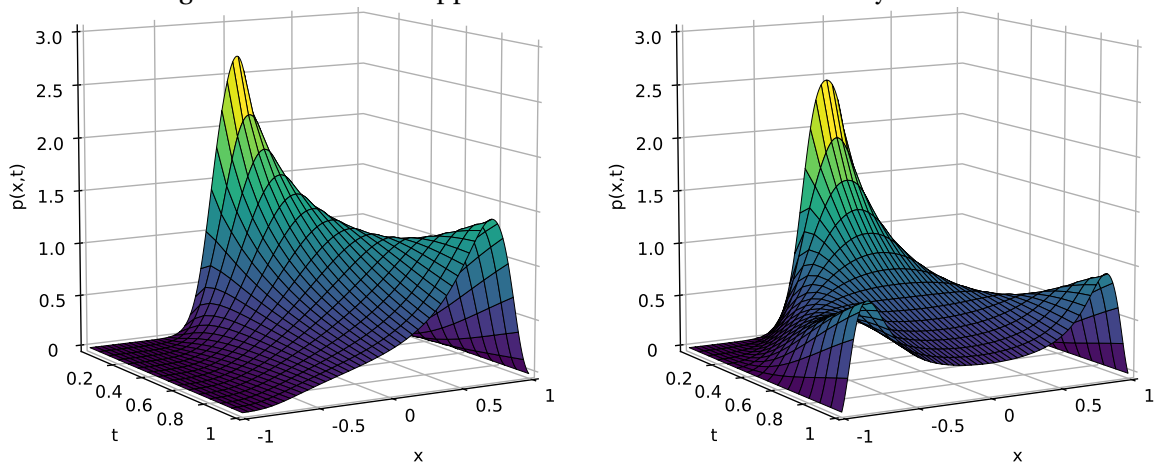


Figure 1.2: Numerical approximation of the transient density function.



1.3 Empirical Data

1.3.1 Sentiment Index Estimation

Business climate surveys and sentiment indices have received increasing attention from academics interested in behavioral finance studies. In addition, microblogging platforms currently host large communities of investors. StockTwits (ST) is the most popular stock-related social media platform, with more than 2 million active users since 2019. ST has implemented a feature that allows users to label their own messages as “Bullish” or “Bearish” to express positive or negative

opinions, respectively, about an asset or the market. Users incorporate "cashtags", a portmanteau of cash+hashtag, composed of a dollar sign and a stock ticker to refer to a particular stock (for instance, $AAPL for Apple). In this study, we focused on 15 stocks among the most discussed financial assets and 5 cryptocurrencies and extracted every message containing the corresponding cashtags. To avoid missing values, we selected only trending stocks that were sufficiently discussed throughout the entire considered sample period.

Our sample period for model estimation is composed of 3 years of data, starting on 2018-01-01 and lasting until 2021-01-01. A total of 38.5% of messages in our sample were already preclassified by users as "Bullish", and 10.0% were preclassified as "Bearish". Table 1.1 provides summary statistics for our data. Every user who mentioned the assets under consideration at least once is included in the database. To compute an estimate of the sentiment index x, we employed sentiment analysis (SA) to classify messages as "Bullish", "Bearish" or "Neutral". We constructed a balanced training dataset of preclassified messages extracted from the ST platform, consisting of approximately 10 million messages classified as "Bullish" and 10 million classified as "Bearish".

Following Oliveira et al. (2016) and Renault (2017), our first step of preprocessing was to edit all control codes[4] to avoid counting them as sentiment values. For the inclusion of a message in the dataset, we imposed a minimum length of one word, not counting edited tags replacing control codes. We preserved punctuation and emojis, as they have been found to increase the precision of classification in previous studies (Renault, 2020; Mahmoudi et al., 2018).

We chose a simple naïve Bayes (NB) classifier, as recent studies have indicated that such classifiers show better performance than more sophisticated and time-consuming algorithms for sentiment analysis (Renault, 2020). We used a bag-of-words model from the sklearn Python library. The NB classifier expresses the relationship between a label y ("Bullish" or "Bearish") and a given dependent feature vector $w_1, ..., w_n$ of words (under the assumption that the features are mutually independent). The probability of a label given a list of words can be expressed as

$$P(y|w_1, ..., w_n) = \frac{P(y)\prod_{i=1}^{n} P(w_i|y)}{P(w_1, ..., w_n)}. \tag{1.13}$$

We collected a sample of 30 million preclassified messages, each of which was labeled as positive or negative by the users. To create a balanced dataset, we used random undersampling to reduce the number of messages to 20 million, with an equal number of positive and negative examples. We then divided the dataset into training and testing sets. To train the classifier, we used k-fold cross-validation with $k = 5$ groups to train the model on the training set. After evaluating the model's performance by comparing the predicted labels to the true labels, we found that it achieved an accuracy of 75%, similar to what has been reported in previous studies (Renault (2017, 2020)).

Once the model was trained, we used it to classify the remaining messages as "Bullish", "Bearish" or "Neutral". The classification results are shown in Table 1.2 for each asset. Prefiltered messages containing no textual information were classified as "Neutral". To assign a sentiment score

[4]All instances of usernames (@john...), URL links (http, https...), hashtags (#business) and cashtags ($AAPL, $MSFT...) were thus substituted with dedicated tags: USERTAG, URLTAG, HASHTAG and CASHTAG.

based on the predicted probabilities, we use a simple thresholding method, as proposed by Renault (2017). For example, if the model outputs a probability of 0.8 for a positive sentiment label, we assign a sentiment score of 0.8 to that sample. To eliminate messages with low probability of classification, we do not consider messages with a probability lower than 20%. Based on this approach, we define the following sentiment measure for a given message, in accordance with its classification probability:

$$S = 2\left(P(y|x_1,...,x_n) - 0.5\right) \tag{1.14}$$

where $S \in [-1,1]$. The next step was to classify users by aggregating the sentiments of their messages into a weekly average[5]. If the weekly average for user i was greater than 0.2, user i was classified as being bullish at time t. If the weekly average was less than -0.2, then user i was classified as bearish. Once the individual sentiments aggregated to the weekly level had been properly classified, we computed the overall sentiment index in the following manner:

$$x_t = \frac{n_+^t - n_-^t}{n_+^t + n_-^t} \tag{1.15}$$

where n_+^t is the total number of individuals with positive sentiments in time interval t and n_-^t is the number of individuals with negative sentiments. We assumed that equal numbers of neutral individuals could be assigned to the optimistic and pessimistic states; then, the resulting sentiment index x_t could be directly used in the model introduced in Section 2. We also assumed that every market participant had the same number of neighbors corresponding to the number of potential interactions measured as the average in-degree of the network. With the networked data at our disposal, we fitted the value D in the model to the average in-degree reported in Table 1.1 of the corresponding users in the interacting network.

[5]Unfortunately, due to limitations in current computational resources, we had to aggregate the data to weekly frequency to reduce the number of observations. The use of daily data would have dramatically increased the computational time of the estimation.

Table 1.1: Social sample summary statistics.

Asset	Users (total)	Posts (total)	Labeled Positive (total)	Labeled Negative (total)	Classified Positive (total)	Classified Negative (total)	In-degree (mean)	Out-degree (mean)
AAPL	86.4	1,184.4	379.0	156.7	494.6	494.1	38.20	282.88
AMD	58.8	1,116.4	468.5	112.3	568.2	347.2	39.72	341.18
AMZN	60.3	827.3	266.6	82.6	369.6	326.6	44.43	428.61
FB	53.3	575.7	169.9	78.6	240.1	243.4	45.95	431.35
GOOG	19.8	100.3	26.0	9.0	47.4	38.3	55.85	819.97
MSFT	43.5	339.4	125.5	22.5	178.5	107.3	44.18	563.71
NFLX	44.7	447.1	121.6	66.1	185.7	190.8	48.60	516.61
NVDA	36.9	331.8	106.9	34.4	160.6	117.8	47.00	560.56
TSLA	117.1	2,207.5	843.9	350.6	878.6	946.1	33.19	251.95
TWTR	30.4	223.1	67.2	23.0	103.5	83.0	53.11	630.89
BIOC	22.5	218.4	110.1	7.4	140.4	41.7	46.74	343.96
FCEL	27.1	403.2	221.3	14.3	255.5	75.2	43.73	533.13
GEVO	18.8	199.3	103.5	7.6	123.6	40.4	53.46	424.22
IBIO	37.6	760.3	432.8	22.9	480.6	135.5	36.53	250.78
XSPA	30.5	548.0	288.4	15.2	364.3	96.4	36.47	421.65
BTC.X	49.5	1,292.3	542.2	157.8	538.8	485.9	40.95	403.03
LTC.X	13.6	173.8	69.8	10.9	89.8	50.1	52.41	478.34
ETH.X	18.9	202.9	77.1	13.2	106.0	59.4	47.75	561.59
TRX.X	9.8	167.1	72.3	9.4	92.5	42.1	49.76	281.72
XRP.X	17.8	192.1	81.2	13.8	98.2	54.0	44.18	354.01

Note: This table shows descriptive statistics for the numbers of users and the numbers of messages (in thousands) posted on the ST platform. The statistics are compiled by asset. Posts could be labeled as either positive (bullish) or negative (bearish) by users and the remaining were automatically classified using the SA methodology. This table also presents the average in-degree and out-degree for each stock, which correspond to the average number of following and the average number of followers per user. These statistics concern the sample period from 2018-01-01 to 2021-01-01.

1.3.2 Stock Market Data

We extracted weekly closing price data from the Bloomberg terminal, where y_t denotes the weekly return at time t. Table 1.2 reports annualized information on asset returns across the sample. The last column provides the correlation between investor sentiment and asset returns. The financial returns and sentiment index for every asset are also visualized in Figures 1.3 and 1.4.

Table 1.2: Stock market data statistics.

Asset	Company Name	Returns (Annualized)	Std. Dev. (Annualized)	Skewness	Excess Kurtosis	$\rho(x, y)$
AAPL	Apple	0.43	0.32	1.17	0.15	0.36
AMD	AMD	0.83	0.54	0.95	-0.15	0.39
AMZN	Amazon.com	0.37	0.30	1.16	0.03	0.41
FB	Facebook	0.19	0.35	0.79	-0.05	0.40
GOOG	Alphabet	0.19	0.26	1.02	0.24	0.38
MSFT	Microsoft	0.35	0.25	0.57	-0.99	0.32
NFLX	Netflix	0.40	0.41	0.63	-0.34	0.33
NVDA	NVIDIA	0.40	0.44	1.22	0.37	0.45
TSLA	Tesla	1.04	0.69	2.07	3.47	0.49
TWTR	Twitter	0.41	0.52	0.76	0.40	0.54
BIOC	Biocept	-0.28	1.60	2.86	9.89	0.35
FCEL	FuelCell Energy	1.51	2.51	0.97	-0.50	0.28
GEVO	Gevo	0.47	1.45	1.85	2.82	0.29
IBIO	iBio	2.23	4.37	1.85	5.69	0.17
XSPA	XpresSpa	-0.42	1.67	2.30	4.91	0.29
BTC-USD	Bitcoin/USD	0.52	0.74	2.01	6.65	0.63
LTC-USD	Litecoin/USD	0.27	0.99	1.81	3.36	0.57
ETH-USD	Ethereum/USD	0.50	1.01	1.80	3.41	0.63
TRX-USD	Tronix/USD	1.19	2.32	4.60	31.66	0.37
XRP-USD	Ripple/USD	-0.10	1.21	4.60	27.86	0.45

Note: This table reports summary statistics for assets as computed from the daily returns. We report the correlation $\rho(x, y)$ between stock returns and investor sentiment at a weekly frequency for 156 observations. These results concern the sample period from 2018-01-01 to 2021-01-01.

Figure 1.3: Weekly financial returns and sentiment index.

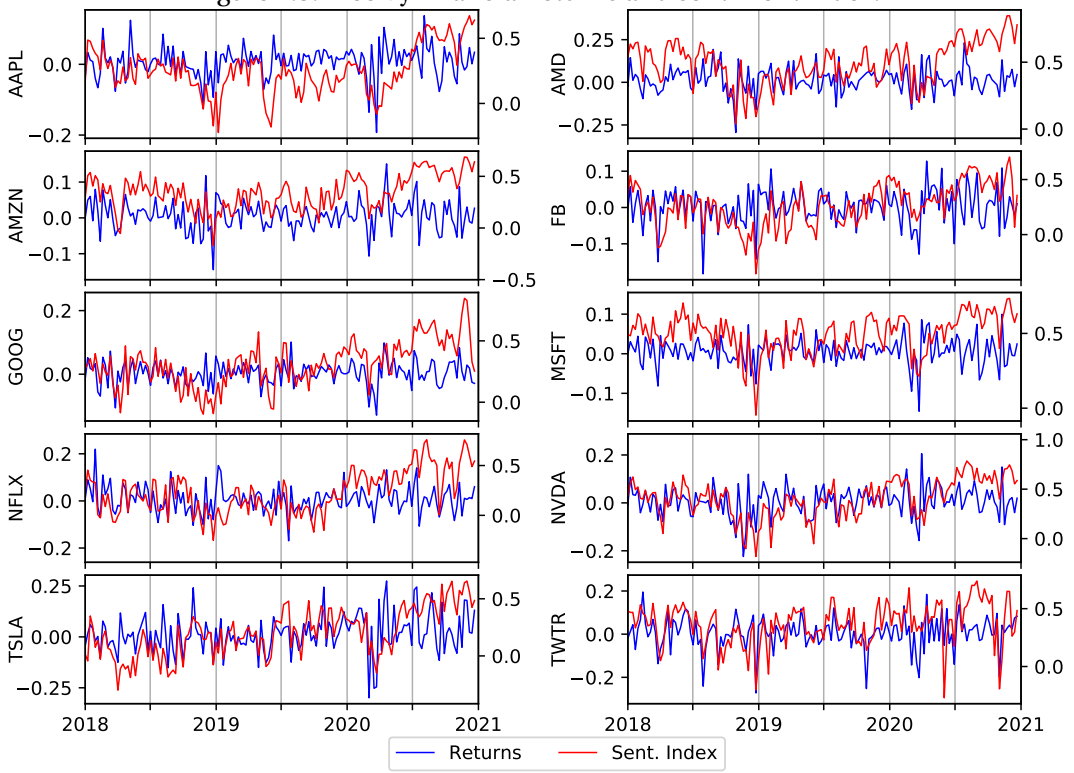
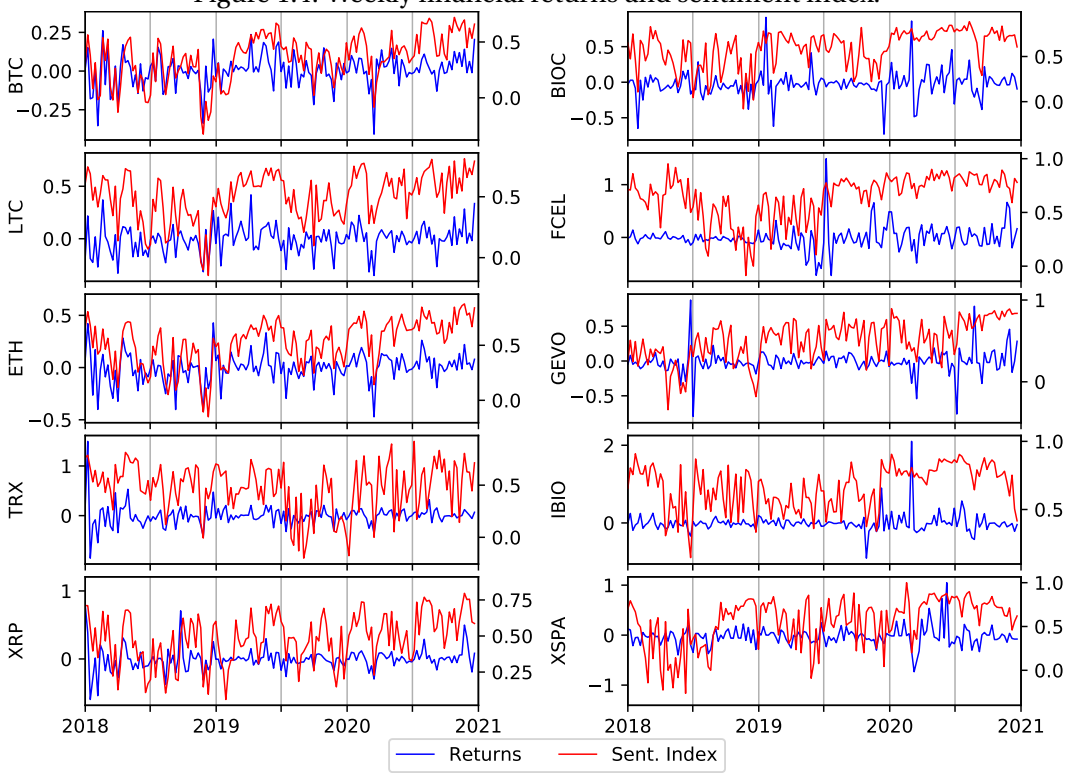


Figure 1.4: Weekly financial returns and sentiment index.



1.4 Empirical Results

1.4.1 Model Estimation

This section presents the results of estimating an ABM on a sample of 156 weekly observations of 20 financial assets during the time period from 2018-01-01 to 2021-01-01. Following our empirical measurement, the parameters were estimated using a social network topology featuring the average in-degree centrality given in Table 1.1. The resulting model provides the ability to estimate the influence of exogenous effects on the opinion formation process. Accordingly, we adapted the influence function $U(\cdot)$ to incorporate asset returns and investor attention. Investor attention on social media is here measured as the scaled number of users posting in the current period. The parameters were estimated in three different frameworks, as follows:

M1: We estimated $U = \alpha_0 + \alpha_1 x_t$, such that the opinion dynamics depend solely on the bias α_0 and the herding parameter α_1.

M2: We modified the model to $U = \alpha_0 + \alpha_1 x_t + \alpha_2 y_t$, where α_2 was added to capture the effect of asset returns y on the opinion formation process.

M3: We further modified the influence function of M2 to obtain $U = \alpha_0 + \alpha_1 x_t + \alpha_2 y_t z_t$, where the effect of financial returns y scales with the level of investor attention z.

The results of the first framework, M1, are shown in Table 1.3 below. The standard errors of the estimated parameters, computed using the negative Hessian matrix evaluated at the maximum likelihood estimator (MLE), are reported in parentheses. The assets are sorted in descending order of the estimates of the herding parameter α_1. These estimates, ranging from 0.49 to 1.08, suggest a strong influence of investor interactions in the formation of sentiment. The α_1 parameter is relatively high for assets that suffered a high level of volatility (reported in Table 1.2), even tending toward 1, which suggests a bimodal density in which predominantly bullish and predominantly bearish states are equally probable. However, BIOC, IBIO and the 5 cryptocurrencies exhibit higher levels of positive bias, indicating that investors might have been more likely to have a fixed bias toward a positive opinion of these assets. The sentiment index is therefore higher for these financial assets.

Table 1.3: Maximum likelihood estimation for M1: $U = \alpha_0 + \alpha_1 x$.

Stock	ν	α_0	α_1	$\log \mathscr{L}$	AIC	BIC
XSPA	1.511 (0.244)	0.060 (0.023)	1.083 (0.043)	-18.957	43.915	53.026
FCEL	0.535 (0.076)	0.079 (0.055)	1.071 (0.089)	-86.73	179.46	188.57
GEVO	0.917 (0.134)	0.072 (0.028)	0.999 (0.056)	-34.516	75.032	84.143
BIOC	0.778 (0.124)	0.122 (0.044)	0.951 (0.078)	-67.033	140.067	149.178
IBIO	0.542 (0.083)	0.187 (0.078)	0.894 (0.123)	-91.894	189.787	198.898
NVDA	0.266 (0.036)	0.002 (0.058)	0.891 (0.130)	-95.742	197.484	206.595
MSFT	0.187 (0.028)	0.003 (0.130)	0.888 (0.248)	-125.297	256.595	265.706
TRX-USD	1.629 (0.309)	0.106 (0.022)	0.878 (0.049)	-24.835	55.67	64.781
ETH-USD	0.458 (0.064)	0.118 (0.048)	0.858 (0.092)	-78.653	163.306	172.417
AMZN	0.215 (0.029)	0.003 (0.063)	0.773 (0.164)	-107.807	221.613	230.724
LTC-USD	0.559 (0.085)	0.160 (0.045)	0.745 (0.090)	-72.58	151.16	160.271
BTC-USD	0.509 (0.073)	0.105 (0.031)	0.695 (0.082)	-64.297	134.594	143.705
NFLX	0.372 (0.049)	0.053 (0.027)	0.652 (0.094)	-72.512	151.025	160.136
AAPL	0.142 (0.018)	0.004 (0.055)	0.576 (0.183)	-134.706	275.412	284.522
TSLA	0.272 (0.036)	0.044 (0.031)	0.576 (0.113)	-93.761	193.522	202.633
XRP-USD	0.491 (0.075)	0.226 (0.052)	0.558 (0.111)	-82.286	170.571	179.682
TWTR	0.724 (0.126)	0.174 (0.033)	0.519 (0.088)	-61.442	128.883	137.994
FB	0.337 (0.046)	0.092 (0.033)	0.517 (0.110)	-85.922	177.845	186.956
GOOG	0.317 (0.043)	0.117 (0.040)	0.514 (0.119)	-89.417	184.834	193.945
AMD	0.212 (0.028)	0.232 (0.085)	0.488 (0.174)	-124.106	254.212	263.323

Note: This table reports the results for M1. The standard errors (reported in parentheses) are computed using the negative Hessian matrix evaluated with the MLE. These results concern 156 observations over the sample period from 2018-01-01 to 2021-01-01.

For the M2 estimates in Table 1.4, the herding intensity α_1 is lower and most often falls in the standard error interval around zero. In most cases, the influence of interactions vanishes when financial returns are considered as an additional explanatory variable. Accordingly, we observe large values for α_2, suggesting that the sentiment process is influenced by the financial returns. It could be that the sentiment dynamics are closely related to the stock prices. However, for assets that suffered high volatility as well as TRX, the herding parameter α_1 remains near 1, while the effect of the financial returns as indicated by α_2 is limited. These results cannot be explained by differences in the correlation coefficient between the returns and the sentiment index (Table 1.2). For instance, we observe similar values of the correlation between returns and sentiment for FCEL ($\rho = 0.28$) and MSFT ($\rho = 0.32$). However, FCEL exhibits the lowest estimated value of α_2 (0.293 +/- 0.069), whereas MSFT exhibits the highest estimated value of α_2 (6.794 +/- 1.256). This observation suggests that investor interaction has a stronger influence on the formation of opinion for highly volatile assets than for other financial assets. We also observe positive values of the bias α_0, consistent with the results for M1.

Table 1.4: Maximum likelihood estimation for M2: $U = \alpha_0 + \alpha_1 x + \alpha_2 y$.

Stock	ν	α_0	α_1	α_2	$\log \mathscr{L}$	AIC	BIC
XSPA	1.455 (0.233)	0.087 (0.024)	1.039 (0.045)	0.293 (0.069)	-30.247	66.493	75.604
FCEL	0.522 (0.076)	0.120 (0.057)	0.996 (0.094)	0.207 (0.067)	-92.554	191.107	200.218
GEVO	0.829 (0.119)	0.092 (0.031)	0.953 (0.061)	0.318 (0.079)	-44.001	94.003	103.114
BIOC	0.722 (0.118)	0.189 (0.049)	0.838 (0.087)	0.402 (0.094)	-79.12	164.24	173.351
IBIO	0.493 (0.073)	0.247 (0.083)	0.794 (0.131)	0.362 (0.041)	-94.203	194.407	203.518
TRX-USD	1.111 (0.189)	0.139 (0.028)	0.789 (0.062)	0.557 (0.106)	-46.449	98.898	108.009
NFLX	0.256 (0.034)	0.047 (0.033)	0.500 (0.117)	3.160 (0.564)	-97.854	201.709	210.82
GOOG	0.216 (0.029)	0.153 (0.049)	0.255 (0.154)	5.622 (0.953)	-117.109	240.219	249.33
XRP-USD	0.315 (0.046)	0.347 (0.073)	0.254 (0.158)	1.236 (0.229)	-115.17	236.34	245.451
FB	0.189 (0.026)	0.113 (0.044)	0.218 (0.154)	5.457 (0.843)	-124.272	254.543	263.654
TWTR	0.429 (0.069)	0.264 (0.046)	0.171 (0.130)	2.383 (0.396)	-95.711	197.422	206.533
LTC-USD	0.284 (0.041)	0.407 (0.073)	0.168 (0.155)	2.070 (0.298)	-123.539	253.078	262.189
NVDA	0.162 (0.022)	0.243 (0.069)	0.149 (0.178)	4.857 (0.734)	-139.244	284.488	293.599
TSLA	0.171 (0.023)	0.051 (0.038)	0.116 (0.160)	2.879 (0.450)	-131.081	268.163	277.274
ETH-USD	0.199 (0.026)	0.446 (0.083)	0.111 (0.168)	2.802 (0.329)	-146.543	299.086	308.197
AMD	0.135 (0.019)	0.332 (0.105)	0.075 (0.227)	3.770 (0.633)	-154.737	315.474	324.585
AAPL	0.092 (0.012)	0.048 (0.063)	0.014 (0.231)	7.751 (1.156)	-170.341	346.682	355.792
AMZN	0.141 (0.019)	0.216 (0.070)	-0.023 (0.207)	7.132 (1.115)	-147.42	300.84	309.951
BTC-USD	0.183 (0.025)	0.305 (0.057)	-0.127 (0.171)	4.172 (0.503)	-141.086	288.172	297.283
MSFT	0.146 (0.021)	0.491 (0.124)	-0.215 (0.259)	6.794 (1.256)	-157.118	320.237	329.347

Note: This table reports the results for M2. The standard errors (reported in parentheses) are computed using the negative Hessian matrix evaluated with the MLE. These results concern 156 observations over the sample period from 2018-01-01 to 2021-01-01.

Table 1.5 displays the results of the third framework. The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) both indicate better performance of M3, which in-

corporates the interaction with investors' attention, over M2. These results suggest an increase in explanatory power when the investor attention component is added. We see strong consistency between the M2 and M3 parameters except for some cases in which the herding parameter becomes significantly higher than 0 but remains relatively small.

Table 1.5: Maximum likelihood estimation for M3: $U = \alpha_0 + \alpha_1 x + \alpha_2 y \times z$.

Stock	ν	α_0	α_1	α_2	$\log \mathscr{L}$	AIC	BIC
XSPA	1.521 (0.246)	0.062 (0.023)	1.076 (0.043)	0.197 (0.109)	-20.959	47.918	57.029
FCEL	0.536 (0.077)	0.084 (0.055)	1.058 (0.089)	0.140 (0.116)	-87.474	180.948	190.059
GEVO	0.910 (0.133)	0.072 (0.029)	0.995 (0.057)	0.169 (0.125)	-35.465	76.93	86.04
BIOC	0.768 (0.122)	0.130 (0.045)	0.930 (0.079)	0.451 (0.195)	-69.868	145.737	154.848
IBIO	0.541 (0.084)	0.195 (0.078)	0.878 (0.124)	0.078 (0.057)	-92.913	191.826	200.937
TRX-USD	1.513 (0.279)	0.111 (0.023)	0.868 (0.051)	1.272 (0.485)	-28.698	63.397	72.508
NFLX	0.302 (0.039)	0.054 (0.030)	0.592 (0.105)	5.998 (1.262)	-86.875	179.751	188.861
LTC	0.448 (0.066)	0.228 (0.052)	0.590 (0.107)	2.550 (0.497)	-90.567	187.133	196.244
ETH-USD	0.325 (0.043)	0.247 (0.061)	0.587 (0.118)	4.715 (0.685)	-110.73	227.46	236.571
XRP-USD	0.434 (0.065)	0.263 (0.057)	0.454 (0.123)	1.817 (0.417)	-93.965	193.931	203.042
TSLA	0.240 (0.032)	0.050 (0.033)	0.378 (0.128)	4.106 (0.886)	-107.493	220.986	230.097
FB	0.265 (0.034)	0.120 (0.038)	0.374 (0.128)	7.072 (1.146)	-105.486	216.972	226.083
GOOG	0.236 (0.031)	0.153 (0.047)	0.327 (0.144)	11.825 (2.153)	-110.728	227.455	236.566
NVDA	0.228 (0.030)	0.240 (0.060)	0.302 (0.151)	6.702 (1.145)	-120.347	246.694	255.804
TWTR	0.494 (0.080)	0.262 (0.043)	0.237 (0.120)	4.227 (0.717)	-89.031	184.062	193.173
AMD	0.177 (0.024)	0.312 (0.094)	0.234 (0.199)	3.855 (0.835)	-138.401	282.801	291.912
BTC-USD	0.284 (0.037)	0.249 (0.047)	0.184 (0.134)	5.259 (0.648)	-116.219	238.437	247.548
AAPL	0.119 (0.015)	0.079 (0.057)	0.179 (0.203)	14.104 (2.653)	-153.332	312.665	321.776
AMZN	0.187 (0.025)	0.215 (0.062)	0.131 (0.179)	8.510 (1.659)	-130.514	267.029	276.14
MSFT	0.180 (0.026)	0.501 (0.115)	-0.126 (0.238)	15.982 (3.617)	-146.402	298.805	307.916

Note: This table reports the results for M3. The standard errors (reported in parentheses) are computed using the negative Hessian matrix evaluated with the MLE. These results concern 156 observations over the sample period from 2018-01-01 to 2021-01-01.

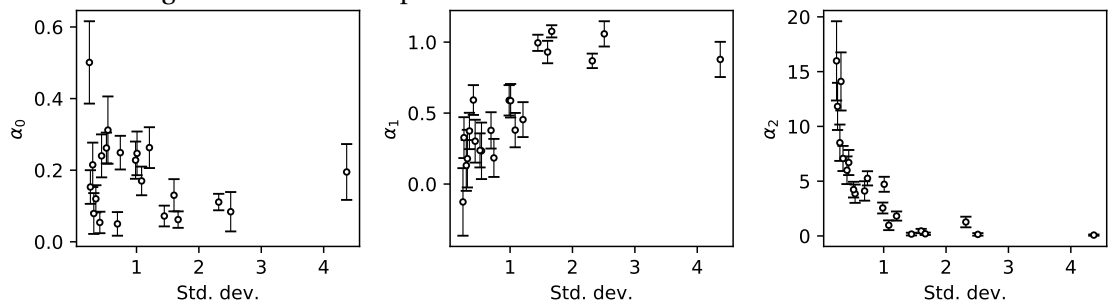
Considering all models, we observe higher values of the parameter ν and a strong herding intensity α_1 for assets such as XSPA and TRX, which are linked to a higher propensity toward sentiment changes. For these assets, switches between extremely bullish and bearish configurations are more likely to occur. This tendency is also coupled with relatively small values of the bias α_0 and of α_2, indicating that investor sentiment is more likely an auto-generated process fed only by investor interactions.

1.4.2 Herding Intensity and Volatility Levels

In previous studies, it has been shown that a strong implication of herding behavior was linked to the generation of stylized facts such as clustered volatility (Lux and Marchesi, 2000; Wang and Wang, 2018; Blasco et al., 2012). In an attempt to explain the differences in the herding estimates, we link the parameters with the volatility level. Figure 1.5 presents a graphical visualization of the

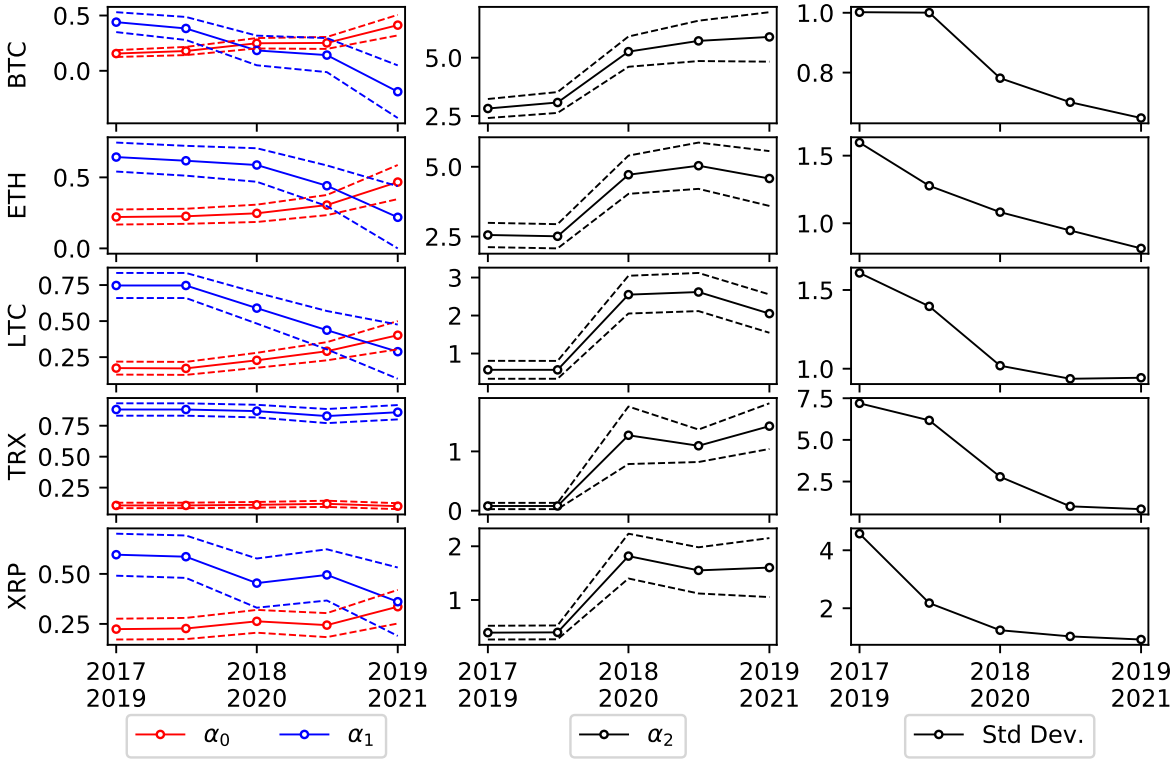
estimated parameters α_0, α_1, and α_2 for M3, where error bars are used to represent the standard errors. The parameter values are plotted against the corresponding standard deviation of the financial assets. We observe that assets with high values of the herding parameter α_1 experienced a stronger level of volatility. The influence of the financial returns, as reflected in the estimated α_2, decreases when assets exhibit higher standard deviations. However, no clear patterns emerge from the analysis of the bias parameter α_0. Overall, these results are consistent with those of Blasco et al. (2012), who show that herding behavior is linked to higher volatility levels.

Figure 1.5: Estimated parameters and the standard deviation of returns.



As stated by Bouri et al. (2019), a static analysis of cryptocurrency markets can be misleading. Therefore, we estimated M3 in a dynamic framework for the five cryptocurrencies. At the end of the 2017 and the beginning of 2018, the cryptocurrency market experienced an outbreak of volatility; this period has been identified as a bubble period. Figure 1.6 presents the results of a dynamic estimation of M3 for the 5 cryptocurrency assets, in which the model was estimated every six months using a moving window of two years. We observe a decrease in the herding intensity after the bubble and a significant increase in the bias toward positive opinions. The herding parameter α_1 was initially approximately 0.5 for BTC, ETH and XRP and 0.75 for LTC and then decreased significantly toward zero. Simultaneously, the process of opinion formation appeared to be increasingly influenced by the bias α_0, which increased significantly, except for TRX, which continued to exhibit the same pattern of a relatively null bias and a high value of the estimated herding parameter. The influence of financial returns and investor attention rose just after the bubble burst, with the effect more than doubling for all 5 cryptocurrencies. In contrast, investors were more likely to ignore financial returns in favor of sentiment-forming interactions during the bubble outbreak. This result is consistent with those of previous studies that show high levels of herding intensity in the new expanding phase of the cryptocurrency market due to the extreme level of volatility (Bouri et al., 2019; Chang et al., 2000).

Figure 1.6: Dynamic estimation of M3 and the standard deviation of returns.



1.5 Discussion and Conclusion

This paper has investigated the explanatory power of an ABM for the opinion formation process for investors' sentiments toward various financial assets. Our framework, which considers a social network topology, overcomes the issue of N-dependence reported in the previous literature. We use SA and text mining techniques to build a weekly aggregation of online messages as a proxy for the sentiment index of investors. Whereas most previous research has emphasized the derivation of a proxy to predict stock returns, we focus on the process underlying the formation of investor sentiment. We find that investor interactions have strong implications for sentiment formation toward volatile stocks, while sentiment toward other financial assets is predominantly explained by financial returns and investor attention. Some assets show evidence of an autogenerated process of sentiment formation in which extreme bullish or bearish configurations are more likely to occur. Finally, we estimate our model in a dynamic framework for the five main cryptocurrencies that experienced a bubble in late 2017. Our results indicate a stronger influence of herding behavior during the bubble period. The effect of investors' interactions subsequently decreased immediately after the bubble burst, and the sentiment process came to be increasingly influenced by a bias toward positive sentiment, financial returns and investors' attention.

Overall, these results have stong implications for the asset pricing literature. In particular, they call into question the validity of the efficient market hypothesis (Fama, 1970) based on the rationality of traders who are supposed to make their decision independently of each other. Financial assets that suffer high-intensity herding behavior will not be priced according to the available

information and thus expose investors to additional risk.

While we have linked herding intensity to the level of volatility, the corresponding implications are not clearly established. In future works, we plan to investigate the herding intensity for a larger number of stocks and link it to stock characteristics such as capitalization, financial ratios, and the macroeconomic environment. This broader analysis will help us to understand where and when the sentiment contagion process occurs. Additionally, while we have assumed here that the contagion process is the same for every investor, recent literature has shown that in fact, opinion dynamics tend to be shaped only by relatively few important nodes in a network (Chen et al., 2021). In particular, the influence of financial "gurus" has been proven by Wang and Wang (2018) to lead to more intensive herding behavior. This result may be stronger in the cryptocurrency market, where, as it has already been observed, even a single individual can initiate bubble-like behavior (Shahzad et al., 2022). This phenomenon could be analyzed by considering different types of investors in our model, such as "opinion leaders", who have higher in-degree centrality in the network and therefore are able to reach more individuals in the contagion process. Finally, our analysis could be conducted with either daily or monthly aggregation to explore whether the estimation results of our model are different when different temporal frequencies are considered.

1.A Appendix

1.A.1 Fokker-Planck Equation

Let $p(n,t)$ be the probability that the investor community has a socioconfiguration of $\{n_+, n_-\}$ at time t. Weidlich (1971) developed a formalization to express the temporal change in the probability distribution $p(n,t)$ via the Fokker-Planck equation. This formalization consists of expressing the transition probabilities for the socioconfiguration in terms of the individuals' transition probabilities as given in (1.2):

$$\begin{aligned} w(n \to (n+1)) \equiv w_\uparrow(n) &= n_- p_+(n) = (N-n)p_+(n) \\ w(n \to (n-1)) \equiv w_\downarrow(n) &= n_+ p_-(n) = (N+n)p_-(n) \end{aligned} \tag{A.1}$$

where $n \to (n+1)$ is equivalent to $\{n_+, n_-\} \to \{n_+ + 1, n_- - 1\}$ and corresponds to a transition from opinion (+) to opinion (-) by one of the individuals with opinion (+). This transition probability can also be expressed in terms of the sentiment index defined in (1.1) rather than in terms of the configuration of the whole system:

$$\begin{aligned} w_\uparrow(n) &= NW_\uparrow(x) = N(1-x)p_+(Nx) \\ w_\downarrow(n) &= NW_\downarrow(x) = N(1+x)p_-(Nx). \end{aligned} \tag{A.2}$$

The transition probabilities for the socioconfiguration (1.2) can be generalized in terms of the following master equation (Weidlich (1971)):

$$\frac{\partial p(n,t)}{\partial t} = -\frac{\partial}{\partial n}\left(\left[w_\uparrow(n) - w_\downarrow(n)\right]p(n,t)\right) + \frac{1}{2}\frac{\partial^2}{\partial n^2}\left(\left[w_\uparrow(n) + w_\downarrow(n)\right]p(n,t)\right) \tag{A.3}$$

where $x = \frac{n}{N}$ and $\Delta x = \frac{\Delta n}{N} = \frac{1}{N}$. By treating x as a continuous variable, we can transform (A.3) into a partial differential equation by expanding the right-hand side as a Taylor series up to terms of the second order. The probability distribution function becomes $P(x,t) = Np(n,t)$ and yields the following Fokker-Planck equation:

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}[A(x)P(x,t)] + \frac{1}{2}\frac{1}{N}\frac{\partial^2}{\partial x^2}[D(x)P(x,t)] \tag{A.4}$$

where $A(x)$ is the drift coefficient, defined as

$$K(x) = W_{\uparrow}(x) - W_{\downarrow}(x) \tag{A.5}$$

and $D(x)$ the fluctuation coefficient, defined as

$$Q(x) = \frac{1}{N}\left[W_{\uparrow}(x) + W_{\downarrow}(x)\right]. \tag{A.6}$$

The same formalization can be applied to $p(\langle n(i)\rangle, t)$, the probability that agent i's neighbors have the socioconfiguration $\{\langle n_+(i)\rangle, \langle n_-(i)\rangle\}$ at time t. The transition probabilities for the socioconfiguration and the sentiment index given in (A.1) and (A.2) now become

$$\begin{aligned} w_{\uparrow}(\langle n(i)\rangle) &= D\tfrac{n_+}{2N}p_+(\langle n(i)\rangle) = D(1-x)p_+(\langle n(i)\rangle) \\ w_{\downarrow}(\langle n(i)\rangle) &= D\tfrac{n_-}{2N}p_-(\langle n(i)\rangle) = D(1+x)p_-(\langle n(i)\rangle) \end{aligned} \tag{A.7}$$

$$\begin{aligned} W_{\uparrow}(x) &= (1-x)p_+(dx) \\ W_{\downarrow}(x) &= (1+x)p_-(dx) \end{aligned} \tag{A.8}$$

where $x = \langle n(i)\rangle / d$ and $P(x,t) = dp(\langle n(i)\rangle, t)$. We recover the probability distribution of the sentiment index $P(x,t)$ in the same manner.

1.A.2 Finite Difference Method

As mentioned by Lux (2009), in such a model of interacting agents, no closed-form solution to the Fokker-Planck equation is usually available. Accordingly, we must rely on numerical approximation techniques. The most common approaches to numerical estimation for solving differential equations are based on the finite difference (FD) method. Lux (2009) showed that for such a model, high-accuracy approximation can be achieved by means of the Crank-Nicolson method combining forward and backward differences. Consider the following Fokker-Planck equation:

$$\frac{\partial f(x,t)}{\partial t} = -\frac{\partial}{\partial x}[A(x;\boldsymbol{\theta})f(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[D(x;\boldsymbol{\theta})f(x,t)] \tag{B.1}$$

where $A(x;\boldsymbol{\theta})$ and $D(x;\boldsymbol{\theta})$ are the drift and diffusion coefficient, of the process and θ is the set of unknown parameters to be estimated. In the FD scheme, both the spatial domain (support of the sentiment index $x \in [-1,1]$) and the time interval are discretized, or divided into a finite number of steps. We consider a spatial grid with a distance step h between $x_j = x_0 + j \cdot h$, where $j = 0, 1, \ldots, N_x$, and time steps of length k from $t = 0$ to the final time T: $t_i = ik$ with $i = 0, ..., N_t$

and $k = \frac{T}{N_t}$. With f_j^i denoting the transient density at (x_j, t_i), partial derivatives are approximated using the Taylor series approach. The following FD approximations are selected for the partial derivatives:

$$\frac{\partial f(x,t)}{\partial t} \approx \frac{f_j^{i+1} - f_j^i}{k} \tag{B.2}$$

$$\frac{\partial f(x,t)}{\partial x} \approx \frac{f_{j+1}^{i+1} - f_{j-1}^{i+1}}{2h} \tag{B.3}$$

$$\frac{\partial^2 f(x,t)}{\partial x^2} \approx \frac{f_{j+1}^{i+1} - 2f_j^{i+1} + f_{j-1}^{i+1}}{h^2}. \tag{B.4}$$

The Crank-Nicolson method consists of taking the average of both the forward and backward difference approximations at intermediate points $\left(i + \frac{1}{2}\right)k$ and $\left(j + \frac{1}{2}\right)h$. For the model equation, the Crank-Nicolson scheme yields

$$\begin{aligned} \frac{f_j^{i+1} - f_j^i}{k} &= \frac{1}{2}\left(\frac{A_{j+1}f_{j+1}^{i+1} - A_{j-1}f_{j-1}^{i+1}}{2h} + \frac{A_{j+1}f_{j+1}^{i} - A_{j-1}f_{j-1}^{i}}{2h}\right) \\ &+ \frac{1}{2}\left(\frac{\frac{1}{2}D_{j+1}f_{j+1}^{i+1} - D_j f_j^{i+1} + \frac{1}{2}D_{j-1}f_{j-1}^{i+1}}{h^2} + \frac{\frac{1}{2}D_{j+1}f_{j+1}^{i} - D_j f_j^{i} + \frac{1}{2}D_{j-1}f_{j-1}^{i}}{h^2}\right) \end{aligned} \tag{B.5}$$

$$\begin{aligned} &\left(-kD_{j-1} - khA_{j-1}\right)f_{j-1}^{i+1} + \left(4h^2 + 2kD_j\right)f_j^{i+1} + \left(-kD_{j+1} + khA_{j+1}\right)f_{j+1}^{i+1} \\ &= \left(kD_{j-1} + khA_{j-1}\right)f_{j-1}^{i} + \left(4h^2 - 2kD_j\right)f_j^{i} + \left(kD_{j+1} - khA_{j+1}\right)f_{j+1}^{i} \end{aligned} \tag{B.6}$$

We rearrange the system into the following form:

$$a_j f_{j-1}^{i+1} + b_j f_j^{i+1} + c_j f_{j+1}^{i+1} = d_j f_{j-1}^{i} + e_j f_j^{i} + f_j f_{j+1}^{i} \tag{B.7}$$

where

$$\begin{aligned} a_j &= -kD_{j-1} - khA_{j-1} \\ b_j &= 4h^2 + 2kD_j \\ c_j &= khA_{j+1} - kD_{j+1} \\ d_j &= kD_{j-1} + khA_{j-1} \\ e_j &= 4h^2 - 2kD_j \\ f_j &= kD_{j+1} - khA_{j+1} \end{aligned} \tag{B.8}$$

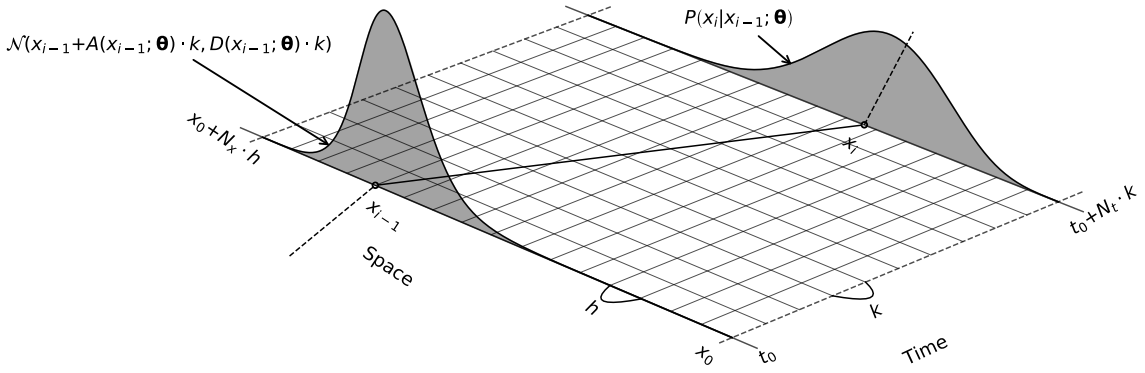
Expressing (B.7) in matrix form, we end up with a computationally convenient tridiagonal system of equations that approximates the continuous-time dynamics of the transient density $f(x,t)$:

$$\mathbf{V}\mathbf{Q}^{i+1} = \mathbf{R}\mathbf{Q}^{i} \tag{B.9}$$

$$\begin{bmatrix} b_0 & c_0 & 0 & 0 & \cdots & 0 \\ a_1 & b_1 & c_1 & 0 & \cdots & 0 \\ 0 & a_2 & b_2 & c_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{N_x-1} & b_{N_x-1} & c_{N_x-1} \\ 0 & \cdots & 0 & 0 & a_{N_x} & b_{N_x} \end{bmatrix} \begin{pmatrix} Q_0^{i+1} \\ Q_1^{i+1} \\ Q_2^{i+1} \\ \vdots \\ Q_{N_x-1}^{i+1} \\ Q_{N_x}^{i+1} \end{pmatrix} = \begin{bmatrix} e_0 & f_0 & 0 & 0 & \cdots & 0 \\ d_1 & e_1 & f_1 & 0 & \cdots & 0 \\ 0 & d_2 & e_2 & f_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & d_{N_x-1} & e_{N_x-1} & f_{N_x-1} \\ 0 & \cdots & 0 & 0 & d_{N_x} & e_{N_x} \end{bmatrix} \begin{pmatrix} Q_0^{i} \\ Q_1^{i} \\ Q_2^{i} \\ \vdots \\ Q_{N_x-1}^{i} \\ Q_{N_x}^{i} \end{pmatrix} \tag{B.10}$$

Coefficients outside the edges of the matrix x_0 and $x_n = x_0 + N_x h$ are equal to zero. The Crank-Nicolson approach requires the assumption that $f(x_0, t) = f(x_n, t) = 0$. Jensen and Poulsen (2002) suggested that the transitional PDF at time $(t_i + k)$ may be well approximated by a normal distribution with a mean of $\mu = x_{i-1} + A(x_{i-1}; \boldsymbol{\theta})\, k$ and a variance of $\sigma^2 = D(x_{i-1}; \boldsymbol{\theta})\, k$. The entire FD procedure must be repeated for each of the N transitions in the dataset, and the likelihood of each transition must be accumulated for the construction of the log-likelihood function. Figure 1.7 presents the discretization procedure for a numerical estimation in which we use $h = 1/16$ and $k = 0.1$ for the discretizations in space and time, respectively, and $T = 1$ as the time horizon. Under the initial conditions, a normal density distribution $\mathcal{N}(x_{i-1} + A(x_{i-1}; \boldsymbol{\theta})k, D(x_{i-1}; \boldsymbol{\theta})k)$ is used to approximate the density distribution.

Figure 1.7: Finite difference scheme.



2

Nonparametric estimator of the tail dependence coefficient: balancing bias and variance

This work has been presented at the following conference:

- ▷ *15th International Conference on Computational and Financial Econometrics (CFE)*, King's College London

Abstract

A theoretical expression is derived for the mean squared error of a nonparametric estimator of the tail dependence coefficient, depending on a threshold that defines which rank delimits the tails of a distribution. We propose a new method to optimally select this threshold. It combines the theoretical mean squared error of the estimator with a parametric estimation of the copula linking observations in the tails. Using simulations, we compare this semiparametric method with other approaches proposed in the literature, including the plateau-finding algorithm.

Keywords: Tail Dependence Coefficient, Nonparametric Estimation, Copula, Extreme Values
JEL Classification: C13, C14, C15, C18

2.1 Introduction

When considering several risk factors, risk managers across various fields such as finance, insurance, hydrology, and engineering, are interested in quantifying the dependence between all these random variables. Although the copula function is the most accurate description of dependence, a simple statistic is often preferred to this function in order to ease the interpretation. Popular examples of such a statistic include quantities based on a linear model, like the Pearson's correlation coefficient, or more realistic nonlinear approaches, such as Spearman's rho. However, these two examples do not specifically focus on the dependence between extreme events, and are thus not relevant in risk management applications. For instance, in finance, stronger dependencies between asset price returns are observed during recessions (Longin and Solnik, 2001; Patton, 2004). Therefore, one might prefer using the tail dependence coefficient (TDC). The TDC depicts the probability that extreme events for several random variables happen simultaneously. It usually refers to the asymptotic probability introduced by Sibuya (1960) and later defined by Joe (1997). It has been used for example in finance (Malevergne and Sornette, 2003; Poon et al., 2004; Caillault and Guégan, 2005) as well as in hydrology, for rainfall data (Poulin et al., 2007; Serinaldi, 2008; Aghakouchak et al., 2010). It is worth noting that the purpose of TDC is not only to determine whether data exhibit tail dependence or not, and hence what type of models might be suitable. There are indeed several applications which require a more accurate estimation of the TDC. For example, in finance, one can base the selection of a portfolio on the TDC, with the motivation of diversifying the portfolio with respect to extreme risks (De Luca and Zuccolotto, 2011).

Before addressing the pivotal question of how to estimate the TDC, it is worth noting that the TDC is a pure copula property: it is not based on marginal distributions but only on the copula, that is the marginal-free version of the joint distribution (Nelsen, 2007; Joe, 2014). Therefore, the estimation of the TDC is strongly related to the estimation of the copula itself.

The parametric estimation of the copula is a first solution that makes it possible to easily derive the TDC. The only challenging step is the choice of the copula function that best fits the data. Such a parametric procedure, in which the whole dataset is used to estimate the copula function, may not be appropriate since it does not focus on the tail. By exploiting extreme value theory, some parametric specifications of the copula however seem natural for depicting the dependence of extreme events (Einmahl et al., 2008; Klüppelberg et al., 2007). This is the case, for example, of the Clayton copula (Juri and Wüthrich, 2002).

To overcome the issue of choosing a specific parameterization of the copula function, some researchers proposed a nonparametric version of the TDC estimator based on the empirical copula introduced by Deheuvels (1979). This estimator corresponds to a discretization of the TDC as defined by Joe (1997) and relies on the selection of a threshold over which the probability of occurrence of joint extreme events is computed. Coles et al. (1999) have motivated a slightly different version of this nonparametric estimator, which is asymptotically equivalent. The selection rule for the threshold strongly impacts the quality of these nonparametric TDC estimators. Ideally, the threshold should make us focus on a few observations only, corresponding to extremes, in order to not bias the TDC estimation with data in the bulk of the distribution. However, the vari-

ance of the estimator would then be overriding. The threshold selection thus corresponds to the art of balancing adequately bias and variance. Most of the existing selection methods are heuristic. Among them, we can cite the plateau-finding algorithm (Frahm et al., 2005b; Schmidt and Stadtmüller, 2006), or graphical methods (Caillault and Guégan, 2005). Most of the contributions in the field are devoted to the comparison of various methods of TDC estimation in simulation frameworks (Frahm et al., 2005b; Schmidt and Stadtmüller, 2006; Poulin et al., 2007; Supper et al., 2020). Yet, to our knowledge, there is no theoretical contribution in which the selection rule of the threshold is related to a simple trade-off between the bias and the variance of the estimator.

We thus propose a theoretical expression for both the bias and the variance of the nonparametric TDC estimator. We then use these expressions to define selection rules in which the threshold in the nonparametric TDC estimator minimizes the theoretical mean squared error (MSE). The formulas depend on the true and unobserved copula. Therefore, a practical application requires choosing a parametric specification for the copula, but only for the tails of the multivariate distribution. To this end, we consider two widespread Archimedean copulas, the Clayton and Gumbel copulas. The Clayton copula offers a flexible representation of tail dependence with various degrees of intensity. Schmidt and Stadtmüller (2006) proposed the Clayton copula to model the tail copula function. Juri and Wüthrich (2002, 2003) showed that the survival Clayton copula is a natural limit for joint excesses beyond a threshold having an Archimedean copula dependence structure. The Gumbel copula is also a natural choice to model upper tail dependence (Galambos, 1978; Joe, 1997) since it is the only copula that is at the same time Archimedean and an extreme-value copula (Genest and Rivest, 1989).

The paper is organized as follows. Section 2 rapidly recalls some basic definitions of the TDC. Section 3 is devoted to theoretical expressions for the bias and variance of nonparametric TDC estimators. Section 4 explores several selection rules for the threshold in the nonparametric TDC estimator. In the simulation study of Section 5, the performance of these estimators is shown to be similar to the one of the plateau-finding algorithm. Section 6 presents a short empirical application to financial data. Section 7 concludes.

2.2 Tail-dependence coefficient

Sklar (1959) showed that any joint distribution of the pair (X, Y) of real random variables can be written as a function of marginal distributions:

$$F(x, y) = C\left(F_X(x), F_Y(y)\right),$$

where C is the copula function between X and Y and can be expressed as:

$$C(u, v) = F\left(F_X^{-1}(u), F_Y^{-1}(v)\right),$$

where $(u, v) \in [0, 1]^2$, F_X^{-1} and F_Y^{-1} are the generalized inverse of the univariate distribution functions F_X and F_Y. The dependence structure is fully described by the copula function and holds independently of the marginal distributions.

In a pioneering article, Sibuya (1960) introduced the notion of tail dependence. This notion describes the dependence between extreme values, either in the upper-right-quadrant tail or in the lower-left-quadrant tail of a bivariate distribution. The lower TDC, denoted λ_L, is defined as follows (Joe, 1997):

$$\lambda_L = \lim_{u \to 0^+} \mathbb{P}\left[X < F_X^{-1}(u) | Y < F_Y^{-1}(u)\right] = \lim_{u \to 0^+} \frac{C(u,u)}{u},$$

if the limit exists. Similarly, the upper TDC is defined by:

$$\lambda_U = \lim_{u \to 1^-} \mathbb{P}\left[X > F_X^{-1}(u) | Y > F_Y^{-1}(u)\right] = \lim_{u \to 1^-} \frac{1-2u+C(u,u)}{1-u},$$

if the limit exists. Since λ_L and λ_U are probabilities, they belong to $[0,1]$.

The tail dependence is a pure copula property, that is, it is independent of the margins of X and Y. The TDC exists if the limits in the above equations exist. If $\lambda_L > 0$ (respectively $\lambda_U > 0$), then the copula presents tail dependence and we simultaneously observe extremely small (resp. extremely large) realizations of X and Y, with conditional probability λ_L (resp. λ_U). In contrast, the absence of tail dependence corresponds to the TDC equal to zero. In this case, the variables X and Y are asymptotically independent.

2.3 Nonparametric estimation of the TDC

The estimation of the TDC is often related to the estimation of copulas. Indeed, if one estimates a parametric copula, one can easily deduce the corresponding parametric TDC. Nonetheless, an accurate estimation of the TDC requires focusing merely on extreme observations[1]. However, the more one confines oneself to extreme observations, the less robust the estimator. The need to rely on extreme observations must thus be balanced with the equally important need to use a sufficiently large amount of data. Nonparametric techniques seem suitable for this purpose, as emphasized by Joe et al. (1992). In this section, we present nonparametric estimators and we introduce their corresponding MSE. We first focus on the lower TDC, then on the upper TDC, and we finish with an extension in which the estimator is itself an average of nonparametric estimators.

2.3.1 Lower tail

We are given n bivariate observations (X_j, Y_j), for $j \in [\![1, n]\!]$, generated with a dependence model of copula C. The nonparametric estimator of the lower TDC is defined by the following (Caillault and Guégan, 2005; Frahm et al., 2005b):

$$\widehat{\lambda}_{L,n}\left(\frac{i}{n}\right) = \frac{\widehat{C}_n\left(\frac{i}{n}, \frac{i}{n}\right)}{\frac{i}{n}}, \tag{1}$$

[1] Otherwise, if one considers that all the observations, even not extreme, are linked by the same copula, other nonparametric methods including all the data are possible. For instance, one may use a nonparametric estimator of Pickands' dependence function, which is an important feature in the definition of an extreme-value copula (Capéraà et al., 1997; Frahm et al., 2005b).

where $(u,v) \in [0,1]^2 \mapsto \widehat{C}_n(u,v)$ is the empirical copula, introduced by Deheuvels (1979). We can write this empirical copula as follows (Genest and Rémillard, 2004):

$$\widehat{C}_n(u,v) = \frac{1}{n}\sum_{j=1}^{n} \{\widehat{F}_{X,n}(X_j) \leq u\}\{\widehat{F}_{Y,n}(Y_j) \leq v\},$$

where $\widehat{F}_{X,n}$ and $\widehat{F}_{Y,n}$ are estimations of the marginal cumulative distribution functions. Focusing on X, $\widehat{F}_{X,n}$ is defined by:

$$\widehat{F}_{X,n}(x) = \frac{1}{n}\sum_{j=1}^{n} \{X_j \leq x\}.$$

Schmidt and Stadtmüller (2006) have shown that the nonparametric estimator of the TDC has a strong consistency and is asymptotically normal.

The estimator of the lower TDC provided in equation (1) relies on the selection of an appropriate integer $i \in [\![1,n]\!]$. Various selection rules for this free parameter have been proposed in the literature. We can cite for example the plateau-finding algorithm (Frahm et al., 2005b) or a graphical method based on monotonic variations of the estimator (Caillault and Guégan, 2005).

Before depicting a new selection criterion for i in equation (1), we have to precise the role of this free parameter in the estimator. The definition of the lower TDC corresponds to the limit case $i/n \to 0$. Nevertheless, as exposed above, using the lowest possible value for i would lead to a non-robust estimator. In contrast, a higher value of i would depict some properties of the copula which are not specifically the ones of its lower tail. Therefore, these two effects, variance and bias, should be balanced in a good compromise. We thus intend to minimize the MSE between the estimator $\widehat{\lambda}_{L,n}(i/n)$ and the true lower tail dependence parameter λ_L. The value of this error is provided in Theorem 2.3.1 in an asymptotic framework.

We introduce some notations that will be used in the theorem. The diagonal section of the copula C is $u \mapsto \delta(u) = C(u,u)$, with the corresponding nonparametric estimator $\widehat{\delta}_n$. The h-function of the copula is the conditional cumulative distribution function provided by $h_1(u,v) = \partial C(u,v)/\partial u$ and $h_2(u,v) = \partial C(u,v)/\partial v$. In the case of a symmetric copula, we will simply write $h(u,v) = h_1(u,v) = h_2(u,v)$. Moreover, the diagonal version of these h-functions, that is when $u = v$, is simplified in $h_1(u)$, $h_2(u)$, and $h(u)$.

Let the bivariate copula C have continuous partial derivatives, $i(n)$ be equal to αn, where $\alpha \in (0,1)$, and $n\left(\widehat{\delta}_n(\alpha) - \delta(\alpha)\right)^2$ be uniformly integrable. Then, the MSE of the nonparametric estimator of the lower TDC, defined in equation (1), behaves asymptotically in the following manner:

$$\left[\left(\widehat{\lambda}_{L,n}\left(\frac{i(n)}{n}\right) - \lambda_L\right)^2\right] = V_{L,n}(\alpha) + \left(\frac{1}{\alpha}\delta(\alpha) - \delta'(0)\right)^2,$$

where

$$\lim_{n\to\infty} nV_{L,n}(\alpha) = \frac{\sigma^2(\alpha)}{\alpha^2}$$

and

$$\sigma^2(\alpha) = \delta(\alpha)(1-\delta(\alpha)) + (1-\alpha)\left[\alpha\left(h_1(\alpha)^2 + h_2(\alpha)^2\right) - 2\delta(\alpha)(h_1(\alpha) + h_2(\alpha))\right] + 2h_1(\alpha)h_2(\alpha)(\delta(\alpha) - \alpha^2). \tag{2}$$

The proof is postponed in Appendix 2.A.1.

The technical condition regarding the uniform integrability of $n\left(\widehat{\delta}_n(i(n)/n)-\delta(i(n)/n)\right)^2$ is fulfilled for example in the case of an independent copula, as exposed in Appendix 2.A.6.

In Theorem 2.3.1, the variance of the estimator is $\sigma^2(\alpha)/n\alpha^2$ and the squared bias is $\left(\delta(\alpha)/\alpha-\delta'(0)\right)^2$. This means that, given $\alpha\in(0,1)$, the variance will shrink to zero as $n\to\infty$ but not the bias. The squared bias and the variance will be more balanced for values of α close to zero and datasets of finite size n, for which we will apply this asymptotic framework. If the copula C is symmetric, $\sigma^2(u)$ more easily writes:

$$\sigma^2(u)=\delta(u)(1-\delta(u))+2(1-u)h(u)\left[uh(u)-2\delta(u)\right]+2h(u)^2(\delta(u)-u^2). \tag{3}$$

We now apply Theorem 2.3.1 to the Clayton copula, which is symmetric:

$$C(u,v)=\left(u^{-\theta}+v^{-\theta}-1\right)^{-1/\theta},$$

where $\theta>0$.

In the case of the Clayton copula of parameter $\theta>0$, the asymptotic variance and squared bias of the nonparametric estimator $\widehat{\lambda}_{L,n}(i(n)/n)$ of the lower TDC, defined in equation (1), with the assumptions of Theorem 2.3.1, are:

$$\begin{cases} \text{variance} &= \frac{1}{n\alpha^2}\left(\delta(\alpha)-\delta(\alpha)^2\left[1+2\frac{2(1-\alpha)(1-\alpha^\theta)+1}{\alpha(2-\alpha^\theta)^2}\right]+2\delta(\alpha)^3\left[\frac{1}{\alpha^2(2-\alpha^\theta)^2}\right]\right)\\ \text{bias}^2 &= \left(\left(2-\alpha^\theta\right)^{-1/\theta}-2^{-1/\theta}\right)^2, \end{cases}$$

where $\delta(\alpha)=\left(2\alpha^{-\theta}-1\right)^{-1/\theta}$. The proof is postponed in Appendix 2.A.2.

2.3.2 Upper tail

We can extend to the upper tail the results exposed above for the lower tail. Starting from the relation between the survival copula $\bar{C}$ and the copula C,

$$\bar{C}(u,v)=u+v-1+C(1-u,1-v),$$

we note that the survival diagonal section is

$$\bar{\delta}(u)=2u-1+\delta(1-u).$$

The upper TDC is the lower TDC of the survival copula (Schmidt and Stadtmüller, 2006). So we have

$$\lambda_U=\bar{\delta}'(0)=2-\delta'(1).$$

Remarking that

$$\delta'(1)=\lim_{t\to1^-}\frac{1-C(t,t)}{1-t}$$

the estimator of the upper TDC naturally follows:

$$\widehat{\lambda}_{U,n}\left(\frac{i}{n}\right) = \frac{1 - 2\frac{i}{n} + \widehat{C}_n\left(\frac{i}{n}, \frac{i}{n}\right)}{1 - \frac{i}{n}}, \tag{4}$$

which also depends on the selection of an appropriate $i \in [\![1, n]\!]$.

Let the bivariate copula C have continuous partial derivatives, $i(n)$ be equal to αn, where $\alpha \in (0,1)$, and $n\left(\widehat{\delta}_n(\alpha) - \delta(\alpha)\right)^2$ be uniformly integrable. Then, the MSE of the nonparametric estimator of the upper TDC, defined in equation (4), behaves asymptotically in the following manner:

$$\left[\left(\widehat{\lambda}_{U,n}\left(\frac{i(n)}{n}\right) - \lambda_U\right)^2\right] = V_{U,n}(\alpha) + \left(\frac{1 - 2\alpha + \delta(\alpha)}{1 - \alpha} - 2 + \delta'(1)\right)^2,$$

where

$$\lim_{n\to\infty} nV_{U,n}(\alpha) = \frac{\sigma^2(\alpha)}{(1-\alpha)^2}$$

and $\sigma^2(\alpha)$ is the same as in Theorem 2.3.1. The proof is postponed in Appendix 2.A.3. Like for Theorem 2.3.1, we can explicitly split the MSE of the nonparametric estimator of the upper TDC, expressed in Theorem 2.3.2, in two components: the variance $\sigma^2(\alpha)/n(1-\alpha)^2$ and the squared bias $\left((1 - 2\alpha + \delta(\alpha))/(1-\alpha) - 2 + \delta'(1)\right)^2$. This result will be useful for values of α close to 1.

We now want to illustrate Theorem 2.3.2 with the particular case of a Gumbel copula, whose expression is:

$$C(u,v) = \exp\left[-\left\{(-\ln(u))^\theta + (-\ln(v))^\theta\right\}^{\frac{1}{\theta}}\right].$$

The MSE of the nonparametric estimator of the upper TDC is then directly related to the parameter θ, as exposed in the following proposition. In the case of the Gumbel copula of parameter $\theta > 1$, the asymptotic variance and squared bias of the nonparametric estimator $\widehat{\lambda}_{U,n}(i(n)/n)$ of the upper TDC, defined in equation (4), with the assumptions of Theorem 2.3.2, are:

$$\begin{cases} \text{variance} &= \frac{1}{n(1-\alpha)^2}\left(\delta(\alpha)[1-\delta(\alpha)] + \delta(\alpha)^2\left[\frac{1}{\alpha} - 1\right]2^{\frac{1}{\theta}}\left[2^{\frac{1}{\theta}-1} - 2\right] + \delta(\alpha)^2 2^{\frac{2}{\theta}-1}\left[\frac{\delta(\alpha)}{\alpha^2} - 1\right]\right) \\ \text{bias}^2 &= \left(\frac{1-2\alpha+\delta(\alpha)}{1-\alpha} - 2 + 2^{1/\theta}\right)^2, \end{cases}$$

for $\delta(\alpha) = \exp\left[-\left\{2(-\ln(\alpha))^\theta\right\}^{\frac{1}{\theta}}\right]$. The proof is postponed in Appendix 2.A.4.

2.3.3 Average of estimators

We are now interested in the average estimator:

$$\widehat{\Lambda}_{L,n}\left(\frac{i_1}{n}, ..., \frac{i_m}{n}\right) = \frac{1}{m}\sum_{k=1}^{m}\widehat{\lambda}_{L,n}\left(\frac{i_k}{n}\right), \tag{5}$$

which is the average of m nonparametric TDC estimators of respective thresholds $i_1, ..., i_m \in [\![1, n]\!]$. Such an average nonparametric estimator appears for example in the plateau-finding algorithm, which we will describe in Section 2.4.1. It is intended to reduce the MSE of the previously introduced estimators. However, many combinations of m isolated estimators are possible, so that the

minimization of the MSE is computationally expensive. For this reason, we put forward a simpler version of this method in the simulation study detailed in Section 2.5, in which $i_1, ..., i_m$ are consecutive numbers. Theorem 2.3.3 provides a formula for the MSE of the average estimator.

Let the bivariate copula C have continuous partial derivatives, $i_k(n)$ be equal to $\alpha_k n$, where $\alpha_k \in (0,1)$, and $n\left(\widehat{\delta}_n(\alpha_k) - \delta(\alpha_k)\right)^2$ be uniformly integrable, for $k \in [\![1, m]\!]$, and $m \geq 1$. Then, the MSE of the average nonparametric estimator of the lower TDC, defined in equation (5), behaves asymptotically in the following manner:

$$\left[\left(\widehat{\Lambda}_{L,n}\left(\frac{i_1(n)}{n}, ..., \frac{i_m(n)}{n}\right) - \lambda_L\right)^2\right] = V_{L,n}^{\Lambda}(\alpha) + \left(\frac{1}{m}\sum_{k=1}^{m}\frac{1}{\alpha_k}\delta(\alpha_k) - \delta'(0)\right)^2,$$

where

$$\lim_{n\to\infty} nV_{L,n}^{\Lambda}(\alpha) = \frac{1}{m^2}\sum_{k,l=1}^{m}\frac{1}{\alpha_k\alpha_l}\mathcal{K}(\alpha_k, \alpha_l)$$

and

$$\begin{aligned}\mathcal{K}(u,v) &= \delta(u\wedge v) - \delta(u)\delta(v) + (h_1(u)h_1(v) + h_2(u)h_2(v))((u\wedge v) - uv)\\ &\quad - h_1(v)(C(u\wedge v, u) - v\delta(u)) - h_2(v)(C(u, u\wedge v) - v\delta(u))\\ &\quad - h_1(u)(C(u\wedge v, v) - u\delta(v)) - h_2(u)(C(v, u\wedge v) - u\delta(v))\\ &\quad + h_1(u)h_2(v)(C(u,v) - uv) + h_1(v)h_2(u)(C(v,u) - uv),\end{aligned}$$

and where $a \wedge b$ is the minimum between a and b. The proof is postponed in Appendix 2.A.5. The extension of this theorem to an average upper TDC estimator is straightforward and is thus omitted.

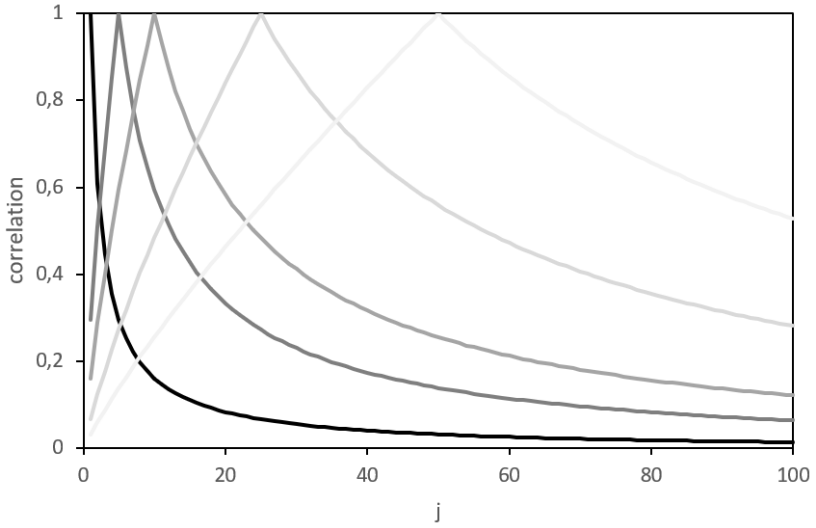
According to Theorem 2.3.3, the variance of the average nonparametric estimator of the lower TDC relies on a function $\mathcal{K}$. If $u = v$, $\mathcal{K}(u,v)$ is simply equal to $\sigma^2(u)$, where $\sigma^2(u)$ is provided by equation 2. The case of C symmetric also simplifies the expression of $\mathcal{K}$:

$$\begin{aligned}\mathcal{K}(u,v) &= \delta(u\wedge v) - \delta(u)\delta(v) + 2h(u)h(v)((u\wedge v) - uv)\\ &\quad - 2h(v)(C(u\wedge v, u) - v\delta(u)) - 2h(u)(C(u\wedge v, v) - u\delta(v))\\ &\quad + 2h(u)h(v)(C(u,v) - uv).\end{aligned}$$

The function $\mathcal{K}$ provides some insights into the dependence between two standard nonparametric estimators. More precisely, if we consider the two estimators of different thresholds, $\widehat{\lambda}_{L,n}(i/n)$ and $\widehat{\lambda}_{L,n}(j/n)$, we can either write their asymptotic covariance as $2\mathcal{K}(i/n, j/n)n/ij$ or as $2\rho_{i,j}\sigma(i/n)\sigma(j/n)n/ij$, where $\rho_{i,j}$ is the correlation between $\widehat{\lambda}_{L,n}(i/n)$ and $\widehat{\lambda}_{L,n}(j/n)$. As a consequence, the asymptotic correlation $\rho_{i,j}$ is equal to $\mathcal{K}\left(i/n, j/n\right)/\sigma\left(i/n\right)\sigma\left(j/n\right)$.

We see in Figure 2.1 this correlation $\rho_{i,j}$ in the case of the Clayton copula. The smaller i, the stronger the correlation decay with respect to j. In other words, the impact on an isolated TDC estimator, when one changes the threshold by a fixed amount, is relatively greater when the initial threshold is extreme. This simply illustrates the lack of statistical robustness of estimators relying on few (extreme) observations: a slight expansion of the data taken into account, with respect to the initial number of observations involved in the estimator, may have significant consequences.

Figure 2.1: Correlation $\rho_{i,j}$ between $\widehat{\lambda}_{L,n}(i/n)$ and $\widehat{\lambda}_{L,n}(j/n)$.



Notes: this figure shows the correlation $\rho_{i,j}$ between $\widehat{\lambda}_{L,n}(i/n)$ and $\widehat{\lambda}_{L,n}(j/n)$ for i equal to (from the darkest to the lightest) 1, 5, 10, 25, and 50, and various values of j. We consider a Clayton copula and $n = 1000$.

2.4 Selection of the threshold

The definition of the nonparametric estimators of the TDC, as in equations (1), (4), and (5), relies on a free parameter i. A proper estimation of the TDC thus requires an appropriate selection of this free parameter, which we subsequently call the *threshold* since the estimators focus on extreme observations whose rank is beyond the threshold i. We first recall a classical selection rule, known as the plateau-finding algorithm (Frahm et al., 2005b). We then propose alternative methods based on the minimization of the asymptotic MSE as expressed in Theorems 2.3.1 and 2.3.2. We finally suggest an extension for average estimators.

2.4.1 Plateau-finding algorithm

Justified by the homogeneity property of the tail copula (Schmidt and Stadtmüller, 2006), the plateau-finding algorithm is a heuristic algorithm that selects the threshold in a characteristic plateau appearing for estimators $\widehat{\lambda}_{L,n}(i/n)$ or $\widehat{\lambda}_{U,n}(i/n)$ of successive i (Frahm et al., 2005b). The algorithm is shown below. Since it works both for the lower and upper tails, we have removed the subscripts L and U. In this paragraph, each $\widehat{\lambda}_n(i/n)$ thus refers to equation (1) or (4).

1. The series $\{\widehat{\lambda}_n(i/n)\}_{i\in[\![1,n]\!]}$ is smoothed using a box kernel with bandwidth $b \in \mathbb{N}$, which consists in applying a moving average on $2b+1$ consecutive elements. We note $\{\bar{\lambda}_n(i/n)\}_{i\in[\![1,n-2b]\!]}$ the new smoothed series, where b is chosen such that 1% of the data falls into the box, that is $b = \lfloor n/200 \rfloor$.

2. We want to select a vector $p_k = \left(\bar{\lambda}_n(k/n), ..., \bar{\lambda}_n((k+m-1)/n)\right)$ of $m = \lfloor\sqrt{n-2b}\rfloor$ consecutive estimates, where $k \in [\![1, n-2b-m+1]\!]$. More precisely, the algorithm selects the index

$k^\star$ of the first[2] vector p_k which satisfies the following plateau condition:

$$\sum_{i=1}^{m-1} \left|\bar{\lambda}_n((k+i)/n) - \bar{\lambda}_n(k/n)\right| \leq 2\sigma,$$

where σ is the standard deviation of the smoothed series $\{\bar{\lambda}_n(i/n)\}_{i\in[\![1,n-2b]\!]}$.

3. Then, the TDC estimator is defined as the average of the estimators $\bar{\lambda}_n(.)$ in the plateau $p_{k^\star}$:

$$\check{\lambda}_n = \frac{1}{m}\sum_{i=0}^{m-1} \bar{\lambda}_n((k^\star+i)/n).$$

If there is no vector fulfilling the plateau condition, the TDC estimate is set to zero.

2.4.2 Minimization of the MSE

As an alternative to the plateau-finding algorithm, we propose selecting the threshold minimizing the asymptotic MSE as expressed in Theorems 2.3.1 or 2.3.2, to balance the bias and the variance of the nonparametric TDC estimator. However, minimizing this MSE leads to two issues. The first is that the formulas of the MSE of the nonparametric estimators depend on the true and unobserved copula of the dataset. We can then imagine a plug-in approach, in which the unobserved copula is replaced in the MSE formula by an empirical estimate. Nonetheless, this leads to the second issue: in addition to the copula itself, the MSE formula includes derivatives of the copula, namely δ' and h. Regarding this ill-posed inverse problem of estimating derivatives, using a simple empirical estimation of the copula is not enough, and regularization is required. We thus propose a parametric specification for the unobserved copula, at least for the plug-in in the MSE formulas of the nonparametric estimators of the TDC.

In this semiparametric approach, we can, for example, assume a Clayton copula when dealing with the lower tail and a Gumbel copula for the upper tail, transforming the MSE formula as in Propositions 2.3.1 and 2.3.2. The method we propose is however more general and one may choose parametric copulas other than these traditional examples of tail-dependent copulas. Whether the copula is a Clayton or a Gumbel, it depends on a parameter θ to be estimated. One could estimate θ using all observations. However, this approach may be strongly biased. We indeed want this specific parametric copula to depict only the tail of the true copula. We thus propose below two competing methods, in which we focus on extreme observations to estimate θ.

We note that this idea of a plug-in to select the most appropriate free parameter of a nonparametric estimator is a very common practice in nonparametric statistics. It is for example widespread in the literature about kernel density estimation (Jones et al., 1996).

[2] Starting from $k = 1$ for the lower TDC and from $k = n - 2b - m + 1$ for the upper TDC.

Simple plug-in approach

We note $\widetilde{MSE}(i, C)$ the MSE provided in Theorem 2.3.1 or in the Theorem 2.3.2. This theoretical asymptotic MSE depends both on the order i used in the nonparametric estimator of the TDC and on the true and unobserved copula C, with respect to which the theoretical MSE is calculated. Since the true copula is unknown, we must estimate it in order to estimate the MSE. As explained above, we consider a model in which the tail of the copula, and only its tail, is close to the tail of a specific parametric copula. For example, one can estimate a Clayton copula by considering only extreme observations, so that the estimate is not influenced by the rest of the dependence structure, which may not be consistent at all with a Clayton copula. This idea is useful for estimating tail copulas and the most widespread method in this perspective is the censored likelihood approach (Smith et al., 1997; Huser and Wadsworth, 2019; Castro-Camilo and Huser, 2020). Here, we propose an estimation method combining the nonparametric estimator and the censored likelihood, both depending on a common threshold which separates the extremes from other observations.

For this purpose, one has to define clearly what are extreme observations. In dimension higher than 1, sorting vectors and thus defining extreme vectors and quantile vectors is a question for which one can conceive several different solutions, such as spatial quantiles (Abdous and Theodorescu, 1992), geometric quantiles (Chaudhuri, 1996), or quantiles based on the inversion of an appropriate mapping (Koltchinskii, 1997; Garcin et al., 2021). Following this last idea, we define here an extreme observation as one belonging to the empirical orthant quantile of probability lower than i/n. In practice, we first determine the probability $\widehat{F}_n$ associated with each observation (X_j, Y_j), that is the empirical probability to have an observation in the lower left orthant of (X_j, Y_j). Then, for a given threshold i, the set of corresponding observations is defined by

$$\Omega^L_{i/n} = \left\{ (X_j, Y_j) \in \mathbb{R}^2, j \in [\![1, n]\!] \,\middle|\, \widehat{F}_n\left(X_j, Y_j\right) \leq \widehat{C}_n\left(\frac{i}{n}, \frac{i}{n}\right) \right\}$$

for the lower tail and

$$\Omega^U_{i/n} = \left\{ (X_j, Y_j) \in \mathbb{R}^2, j \in [\![1, n]\!] \,\middle|\, \widehat{F}_n\left(X_j, Y_j\right) \geq \widehat{C}_n\left(\frac{i}{n}, \frac{i}{n}\right) \right\}$$

for the upper tail.

We note that the probability for a pair of observations to be in $\Omega^L_{i/n}$ is $K(C(i/n, i/n))$, where K is the Kendall function associated with the probability distribution F: $K(p) = [F(X, Y) \leq p]$. We justify this Kendall quantile approach by the fact that a vector (X_j, Y_j) dominates all the observations whose probability is lower than $\widehat{F}_n(X_j, Y_j)$ and not only those in the lower left quadrant of (X_j, Y_j) Garcin et al. (2021). The Kendall function will be overriding for calculating censored likelihoods. It is worth noting that the Kendall function is unique for a given copula and that its expression is straightforward in the case of Archimedean copulas, namely it is $K_\theta(p) = p - p\ln(p)/\theta$ for the Gumbel copula and $K_\theta(p) = p + p^2(1 - p^\theta)/\theta$ for the Clayton copula (Garcin et al., 2021).

Given a threshold i, one estimates a parametric copula close to the true copula of the extreme vectors by a censored maximum likelihood method, restricted to the observations either in $\Omega^L_{i/n}$

or in $\Omega^U_{i/n}$. We note c_θ and C_θ the parametric copula density and cumulative distribution function, which are not specified more precisely here[3] and which are parameterized by θ. We also note K_θ the parametric Kendall function corresponding to the copula C_θ. In this censored approach, the considered likelihood is $c_\theta(\widehat{F}_{X,n}(X_j), \widehat{F}_{Y,n}(Y_j))$ for any vector in the set of extreme observations $\Omega^L_{i/n}$, whereas we replace this likelihood by the probability measure of the set $\mathbb{R}^2 \setminus \Omega^L_{i/n}$ for any non-extreme observation. In order to take into account the parameter θ in this last probability, we consider a pseudo probability, where only the Kendall function depends on θ. Therefore, with this assumption, the probability measure of $\mathbb{R}^2 \setminus \Omega^L_{i/n}$ is $1 - K_\theta(\widehat{F}_n(X_j, Y_j))$, where (X_j, Y_j) is on the boundary of the set $\Omega^L_{i/n}$. The estimator $\widehat{\theta}_{L,i/n}$ of θ is thus, for the lower tail:

$$\widehat{\theta}_{L,i/n} = \underset{\theta}{\operatorname{argmax}} \sum_{j=1}^{n} \left\{ \ln\left(c_\theta(\widehat{F}_{X,n}(X_j), \widehat{F}_{Y,n}(Y_j))\right) \mathscr{J}^L_{j,i} + \ln\left(1 - K_\theta\left(\widehat{C}_n\left(\frac{i}{n}, \frac{i}{n}\right)\right)\right)\left(1 - \mathscr{J}^L_{j,i}\right) \right\},$$

where $\mathscr{J}^L_{j,i} = {}_{(X_j,Y_j)\in\Omega^L_{i/n}}$. We have a similar formula for the upper tail:

$$\widehat{\theta}_{U,i/n} = \underset{\theta}{\operatorname{argmax}} \sum_{j=1}^{n} \left\{ \ln\left(c_\theta(\widehat{F}_{X,n}(X_j), \widehat{F}_{Y,n}(Y_j))\right) \mathscr{J}^U_{j,i} + \ln\left(K_\theta\left(\widehat{C}_n\left(\frac{i}{n}, \frac{i}{n}\right)\right)\right)\left(1 - \mathscr{J}^U_{j,i}\right) \right\},$$

where $\mathscr{J}^U_{j,i} = {}_{(X_j,Y_j)\in\Omega^U_{i/n}}$.

In both formulas, the likelihood is in fact a pseudo likelihood insofar as it uses the empirical marginal distributions instead of a parametric specification with parameters to be estimated. In the simulation study, we even work directly with pseudo observations, insofar as we simulate random variables with a uniform marginal distribution. This approach is a common practice for estimating copulas and leads to an estimator of copula parameters which is asymptotically normal and consistent (Genest et al., 1995; Shih and Louis, 1995). Apart from this consideration, the estimates $\widehat{\theta}_{L,i/n}$ and $\widehat{\theta}_{U,i/n}$ depend on the choice of the threshold i and we can naturally define a mapping ψ such that $\widehat{\theta}_{L,i/n} = \psi(i/n)$, or $\widehat{\theta}_{U,i/n} = \psi(i/n)$ if we are instead interested in the upper tail.

Using a plug-in approach, the MSE is now estimated by $\widehat{MSE}(i, C_{\psi(i/n)})$, which can be expressed thanks to Proposition 2.3.1 (respectively Proposition 2.3.2) if we use the Clayton (resp. Gumbel) specification for lower (resp. upper) tails. The optimal i in this plug-in approach is then:

$$i^\star_{PI} = \underset{i\in[\![1,n]\!]}{\operatorname{argmin}}\ \widehat{MSE}(i, C_{\psi(i/n)}).$$

In other words, given a threshold i, we calculate the corresponding estimated MSE between the nonparametric TDC using this threshold and a TDC for a parametric copula, which we estimate on observations beyond the same threshold i. Finally, we select the threshold $i^\star_{PI}$ minimizing this MSE. The TDC estimator is then the standard nonparametric TDC estimator for the selected

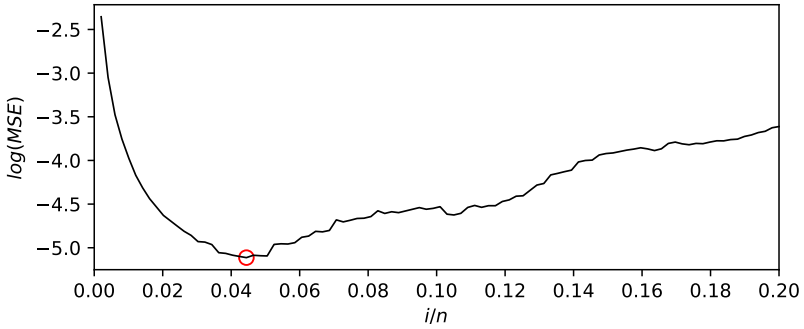
[3] As previously explained, this copula may for example be a Clayton copula for the lower tail or a Gumbel copula for the upper tail. We can even work with a set of various parametric copulas and finally select the pair of copula specification and parameter with the highest likelihood or the highest AIC/BIC. This extension, though promising, is not developed further in this paper. In particular, it implies an explicit formula for the MSE of each specification of parametric copula, as we do for Clayton and Gumbel copulas.

threshold:

$$\widehat{\lambda}_{L,MSE,n} = \widehat{\lambda}_{L,n}\left(\frac{i^{\star}_{PI}}{n}\right),$$

where L has to be replaced by U for the upper case. Figure 2.2 illustrates the selection of this threshold in the case of a lower tail, with data simulated by a rotated Gumbel copula.

Figure 2.2: Function of $\log\big(\widetilde{MSE}(i, C_{\psi(i/n)})\big)$.



Notes: This figure shows the function of $\log\big(\widetilde{MSE}(i, C_{\psi(i/n)})\big)$ with respect to the threshold i/n, for $n = 1000$ pairs simulated with a rotated Gumbel copula of parameter 1.5.

Two-step plug-in approach

In the simple plug-in approach described above, we have selected the rank $i^{\star}_{PI}$ minimizing the estimated MSE, either for the lower or for the upper TDC. However, if we consider that $C_{\psi(i^{\star}_{PI}/n)}$ is the best estimation of the true copula C, at least in the tail of the distribution, one could then argue that the rank $i^{\star}_{PI}$ is not necessarily the one minimizing the estimated MSE of the TDC estimator: in other words, one may find a rank i such that $\widetilde{MSE}(i, C_{\psi(i^{\star}_{PI}/n)}) < \widetilde{MSE}(i^{\star}_{PI}, C_{\psi(i^{\star}_{PI}/n)})$. We are thus eager to find a rank i such that it minimizes the MSE of the TDC estimator with the copula $C_{\psi(i/n)}$: this rank i must verify the following fixed-point equation:

$$i = \underset{j\in[\![1,n]\!]}{\operatorname{argmin}}\ \widetilde{MSE}(j, C_{\psi(i/n)}). \tag{6}$$

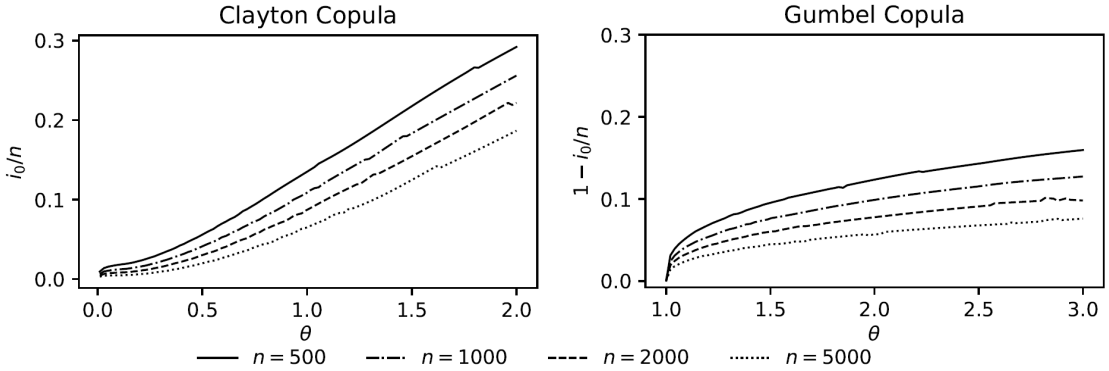
This objective leads to a two-step plug-in approach described below.

From Theorem 2.3.1, given a copula C_θ, we can determine the threshold i_0 minimizing the theoretical asymptotic MSE of the nonparametric TDC estimator. We can thus define a mapping ϕ between the parameter θ and the corresponding optimal threshold in the nonparametric TDC estimator: $i_0/n = \phi(\theta)$. The formal definition of ϕ is as follows:

$$\phi(\theta) = \frac{1}{n}\ \underset{j\in[\![1,n]\!]}{\operatorname{argmin}}\ \widetilde{MSE}(j, C_\theta). \tag{7}$$

If we plot this function ϕ in the particular case of a Clayton or a Gumbel copula, we observe a strictly monotonic function, as one can see in Figure 2.3. In these cases, we can numerically invert ϕ. In a broader perspective, we can define the generalized inverse function, $\phi^{-1} : u \in [0,1] \mapsto \inf\{\theta \in \mathbb{R}, \phi(\theta) \geq u\}$, and finally write $\theta = \phi^{-1}(i_0/n)$.

Figure 2.3: Optimal threshold $\phi(\theta)$.



Notes: this figure represents the optimal threshold $\phi(\theta)$ (respectively $1-\phi(\theta)$) of the lower (resp. upper) TDC estimator, given the copula parameter θ, for various values of n, for the Clayton (resp. Gumbel) copula.

The θ parameter being unknown, we now replace it by its estimator $\widehat{\theta}_{L,i/n}$. As stated in equation (6), we are looking for a rank i defining the same threshold for the estimation of θ and for the estimation of the TDC. Therefore, combining equations (6) and (7) leads to the following equation for the optimal i:

$$\frac{i}{n} = \frac{1}{n} \underset{j \in [\![1,n]\!]}{\operatorname{argmin}} \widetilde{MSE}(j, C_{\psi(i/n)}) = \phi\left(\psi\left(\frac{i}{n}\right)\right).$$

In other words, our objective is to have

$$\psi\left(\frac{i}{n}\right) = \phi^{-1}\left(\frac{i}{n}\right). \tag{8}$$

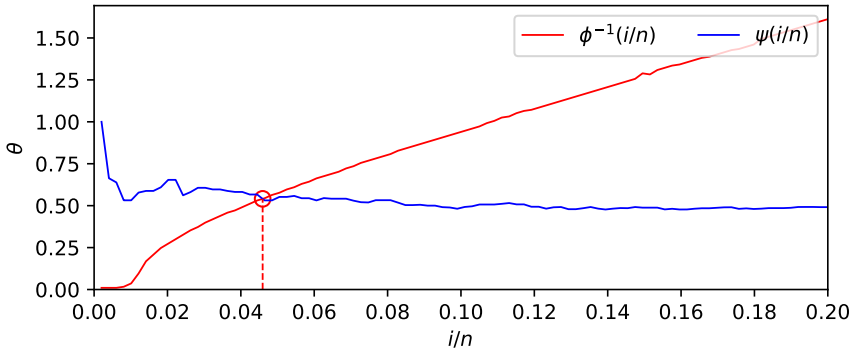
We thus define the optimal threshold i as the rank for which the graphs of ψ and ϕ^{-1} intersect themselves. Depending on the observations and on the parametric copula, this intersection may not exist or be multiple. Moreover, the discrete nature of the threshold makes highly improbable the existence of a fixed point, that is of a i satisfying equation (8). Therefore, we instead minimize the quadratic deviation between $\psi(i/n)$ and $\phi^{-1}(i/n)$, so that our estimated optimal threshold $i^{\star}_{2PI}$ is such that:

$$i^{\star}_{2PI} = \underset{i \in [\![1,n]\!]}{\operatorname{argmin}} \left(\psi\left(\frac{i}{n}\right) - \phi^{-1}\left(\frac{i}{n}\right)\right)^2.$$

In practice, the determination of $i^{\star}_{2PI}$ thus amounts to an optimization algorithm. We can for instance propose to use Nelder-Mead's algorithm (Nelder and Mead, 1965). In this case, we start with two different and arbitrary thresholds, for which we determine the output of the objective function $i \mapsto \left(\psi(i/n) - \phi^{-1}(i/n)\right)^2$. Then, the iteration rule of Nelder-Mead makes these two thresholds evolve and finally converge towards a local minimum. Refinements, such as the mix of several executions of this algorithm, could improve the results and lead to reach a threshold closer to a global minimum. This heuristic approach provides satisfying results in simulations. Figure 2.4 illustrates the principle of the selection of the optimal threshold, corresponding to the abscissa of the intersection of the two curves. In this example, the optimal threshold is lower with the two-step plug-in estimator than with the simple plug-in illustrated in Figure 2.3.

Similarly to the simple plug-in approach, the TDC estimator in this two-step plug-in is the

Figure 2.4: Threshold selection for the two-step plug-in approach.



Notes: threshold selection for the two-step plug-in approach for the lower TDC. The number n of simulated pairs is equal to 1000, and the data-generating copula is a rotated Gumbel copula of parameter 1.5. The θ at the ordinate is the parameter of the Clayton copula used by our estimator to describe the lower tail dependence.

nonparametric TDC estimator for the selected threshold:

$$\widehat{\lambda}_{L,MSE2,n} = \widehat{\lambda}_{L,n}\left(\frac{i^{\star}_{2PI}}{n}\right),$$

where L has to be replaced by U for the upper tail.

2.4.3 Minimizing an average MSE

The plateau algorithm leads to estimations of the nonparametric TDC with a particularly low variance. We can explain this characteristic of this estimation method by the double regularization of the TDC which is performed in steps 1 and 3 of the algorithm detailed in Section 2.4.1. Inspired by this regularization in the plateau algorithm, we can also propose an additional regularization of the MSE-based plug-in estimators.

The smoothing procedure we propose is similar to step 3 of the plateau algorithm. We do not mimic step 1, which consists of first smoothing the estimated nonparametric TDC, because this step would modify the distribution of this TDC and make our MSE estimate erroneous.

We note that the smoothing used in the plateau is a simple rule-of-thumb averaging. It could be beneficial to use other smoothing techniques based on the minimization of the error induced by smoothing. Among these techniques, which are omnipresent in nonparametric statistics (Härdle et al., 2012), one can cite smoothing with wavelets (Mallat, 1999; Ranta, 2010; Garcin and Guégan, 2016; Garcin and Goulet, 2019) or smoothing resulting from a variational problem (Garcin, 2017). We will not use these methods here, to be congruent with the plateau algorithm and to make fair comparisons between the various TDC estimators. Nevertheless, we will see that beyond the arbitrary averaging, one can also do an averaging minimizing the MSE, as a result of Theorem 2.3.3.

We are given a parameter m describing the size of an interval $\mathscr{I}_m$ of consecutive ranks, where m is the size of the plateau in the algorithm described in Section 2.4.1. We now have to select an

appropriate interval $\mathscr{I}_m$ of consecutive ranks, so that our new regularized TDC estimator will be:

$$\widehat{\lambda}_{L,\mathscr{I}_m,n} = \frac{1}{m} \sum_{i \in \mathscr{I}_m} \widehat{\lambda}_{L,n}\left(\frac{i}{n}\right),$$

where L has to be replaced by U for the upper tail. We can propose several ways of selecting $\mathscr{I}_m$. For example, from the knowledge of an optimal plug-in rank, $i^\star_{PI}$ or $i^\star_{2PI}$, we can build $\mathscr{I}_m$ as an interval having this optimal rank in its median or in one of its bounds, such as $[\![i^\star_{2PI}, i^\star_{2PI} + m - 1]\!]$.

Alternatively, we can select an interval of ranks minimizing the average of the MSE of each rank. More precisely, for each rank i, we are able to approximate the corresponding MSE of the nonparametric TDC estimator, following the simple plug-in approach: $\widetilde{MSE}(i, C_{\psi(i/n)})$. We now select the m consecutive ranks leading to the minimal average estimated MSE. We note $k^\star_{PI}$ the left bound of this interval of m ranks:

$$k^\star_{PI} = \underset{k \in [\![1, n-m+1]\!]}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \widetilde{MSE}(k+i-1, C_{\psi((k+i-1)/n)}).$$

The resulting estimated TDC is the average of the TDC estimates whose rank is in the interval $\mathscr{I}_m = [\![k^\star_{PI}, k^\star_{PI} + m - 1]\!]$. Finally, we can also select $\mathscr{I}_m$ using Theorem 2.3.3. Indeed, given $\mathscr{I}_m$, this theorem makes it possible to calculate directly the MSE of the average estimator instead of the average MSE of isolated estimators. We can thus extend naturally the direct and two-step plug-in approaches, in which we replace the mappings ψ and ϕ respectively by Ψ and Φ, which are defined as follows. Given an interval of ranks $\mathscr{I}_m$, the mapping Ψ provides an average estimator of the copula parameter, by focusing on each extreme set $\Omega^L_{i/n}$ corresponding to each rank i of the interval $\mathscr{I}_m$:

$$\Psi(\mathscr{I}_m) = \frac{1}{m} \sum_{i \in \mathscr{I}_m} \psi\left(\frac{i}{n}\right).$$

Given a copula parameter θ, Φ provides the interval $\mathscr{I}_m = [\![i^\star, i^\star + m - 1]\!]$ minimizing the MSE of an average estimator $\widehat{\Lambda}_{L,n}$ of the TDC:

$$i^\star = \underset{i \in [\![1, n-m+1]\!]}{\operatorname{argmin}} \left[\left(\widehat{\Lambda}_{L,n}\left(\frac{i}{n}, ..., \frac{i+m-1}{n}\right) - \lambda_L\right)^2\right].$$

This MSE is expressed in Theorem 2.3.3, which can be applied for example with a Clayton copula.

2.5 A simulation study

We compare the estimators introduced above with other common TDC estimators. More precisely, the four estimators based on the minimization of an MSE include a plug-in approach in which we estimate θ on the whole dataset (as evoked in the preamble of Section 2.4.2), along with other approaches in which we estimate θ using only tail data, which include the simple plug-in (Section 2.4.2), the two-step plug-in (Section 2.4.2), and a simple plug-in average estimator (Section 2.4.3). We focus our analysis on the upper TDC, where the tail dependence function is given by the Gumbel copula.

We compute the empirical bias and standard deviation $\sigma(\widehat{\lambda}_{U,n})$ of the estimator for $N = 100$ random sample replications of three different sample sizes $n \in \{500, 2000, 5000\}$. We also compute the root-mean-square error (RMSE) of the estimator to analyse the trade-off between bias and variance for all estimation methods:

$$\text{RMSE}\left(\widehat{\lambda}_{U,n}\right) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\widehat{\lambda}_{U,n}^{j} - \lambda_U\right)^2}.$$

We compare the estimators using random data of different samples generated by four different bivariate distributions. We first use a Gumbel copula with parameter values $\theta \in \{1.1, 1.5, 1.75, 2\}$, corresponding to $\lambda_U \in \{0.12, 0.41, 0.51, 0.59\}$. The second generated distribution is a bivariate standard t-distribution with $\nu \in \{1, 2, 3\}$ degrees of freedom for correlation values $\rho \in \{0, 0.25\}$, which correspond to six possible TDCs, $\lambda_U \in \{0.29, 0.18, 0.12, 0.39, 0.27, 0.20\}$. Third, we generate a distribution with a survival Clayton copula with parameter values $\theta \in \{0.1, 0.5, 1, 1.5\}$, corresponding to $\lambda_U \in \{0, 0.25, 0.50, 0.63\}$. Finally, we focus on a case with tail independence ($\lambda_U = 0$) corresponding to a Gaussian distribution with correlations $\rho \in \{0, 0.25, 0.5, 0.75\}$.

For convenience, Table 2.1 reports the identification number for each of the eight estimators implemented in this study. The arbitrary threshold selection (1) and (2) serve as baseline indicators. We implement also the maximum likelihood estimator (3), The plateau-finding algorithm (4), whereas the other estimators (5) and (6) are the proposed methods based on minimization of the theoretical MSE.

Table 2.1: Estimation methods.

Method	Description
(1)	Arbitrary choice of the threshold = 1%
(2)	Arbitrary choice of the threshold = 2%
(3)	Maximum likelihood estimation with an arbitrary copula function (Gumbel)
(4)	Plateau-finding algorithm
(5)	Minimization of the MSE: Simple plug-in estimator
(6)	Minimization of the MSE: Two-step plug-in estimator

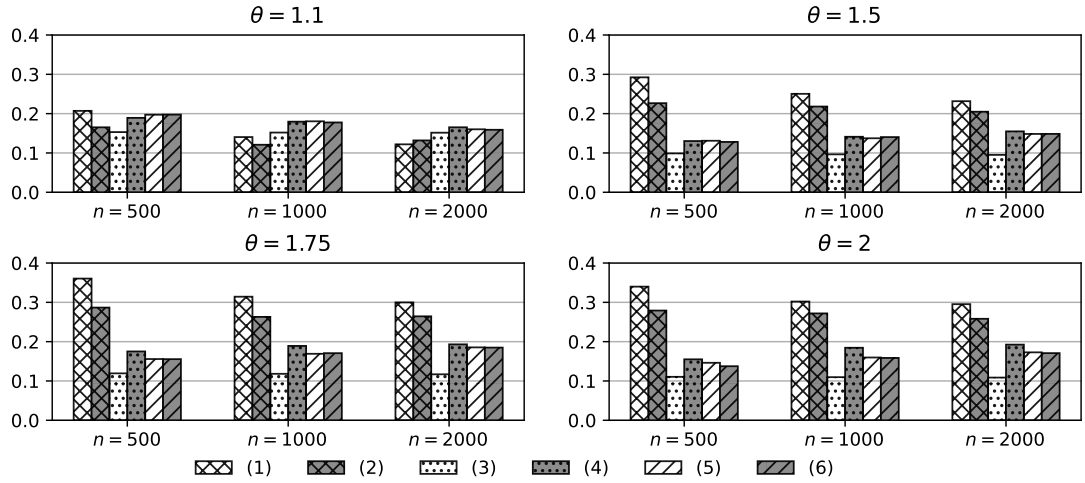
2.5.1 Gumbel simulations

In this case, we assume that the underlying distribution function is known. For estimator (3) we use a maximum likelihood estimation of the Gumbel copula distribution which is the true sample distribution. For the two estimators introduced in this paper (5-6), the function of the upper TDC estimator is also based on the Gumbel distribution.

The results are gathered in Figure 2.5, with more details in Table 2.4 , in the appendix. For the lowest value of the true TDC ($\lambda_U = 0.12$), that is, when the Gumbel copula has a parameter $\theta = 1.1$, all the estimators show almost similar results in terms of RMSE. In this case, even if the methods (1) and (2) exhibit the lowest bias they have the highest variances. However, for the three other datasets, that is for $\theta \in \{1.5, 1.75, 2\}$, when the true TDC is higher, the methods (1) and (2) are the worst performing estimators in terms of bias, variance and RMSE.

The plateau-finding algorithm (4) and the two proposed methods (5) and (6) have good performance, with a slightly lower bias and variance overall for methods (5) and (6). The method relying on the maximum likelihood estimation (3) is the best performing estimator overall: not surprisingly, the estimator based on the estimation of the true copula performs better than the others.

Figure 2.5: RMSE for the upper tail dependence with 100 Gumbel simulations.



2.5.2 Student simulations

Following Schmidt and Stadtmüller (2006), we test our estimator on random generations from a bivariate standard t-distribution with $\nu = 1,2,3$ degrees of freedom. We consider the case with no correlation, $\rho = 0$, and the case where there is a small correlation coefficient, $\rho = 0.25$.

The results are gathered in Figure 2.6, with more details in Tables 2.5 and 2.6 , in the appendix. Excepted for estimator (3), the larger the sample size is, the lower the RMSE. Consistent with the findings of Schmidt and Stadtmüller (2006), the plateau algorithm (4) performs well regardless of the parameters of the generating model. However we observe quite similar results for estimators (5) and (6) (minimization of the MSE).

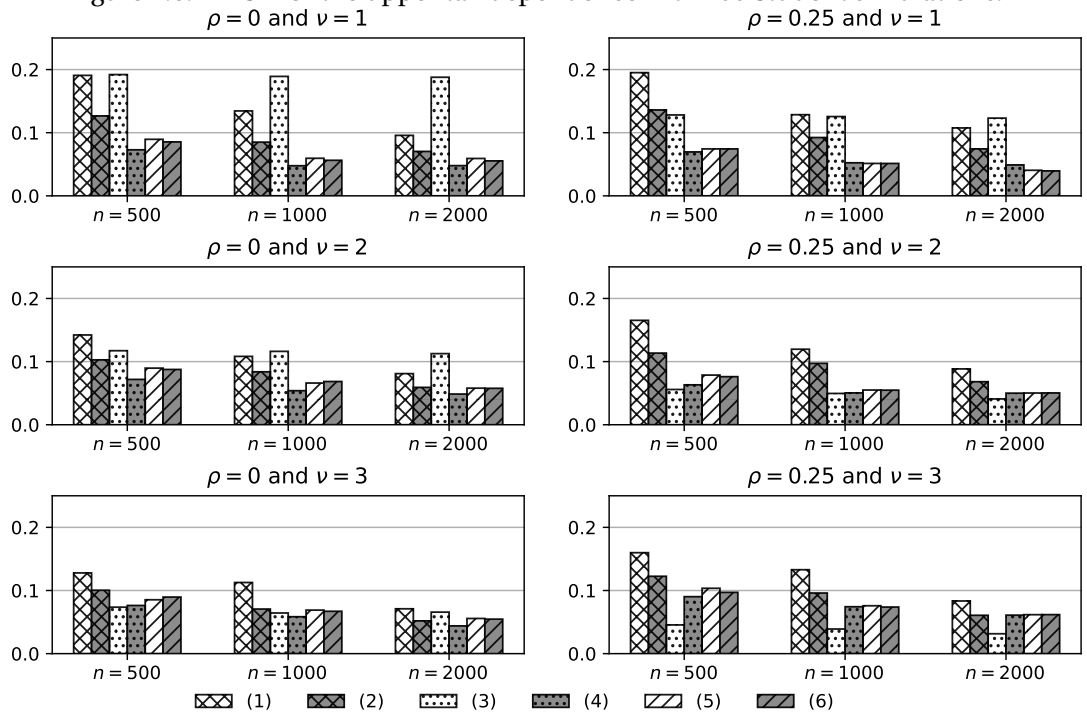
For the method based on the maximum likelihood (3) the performance is strongly dependent on the parameterizations considered. It is the best performing estimator for $\rho = 0.25$ and $\nu \in \{2,3\}$, but it is the worst performing estimator when $\rho = 0$ and $\nu \in \{1,2\}$. The performance for the first two estimators (1) and (2) are also quite dependent on the dataset considered and are not performing well overall, compared to the other estimators.

2.5.3 Rotated Clayton simulations

For these simulations, the results are gathered in Figure 2.7, with more details in Table 2.7 , in the appendix.

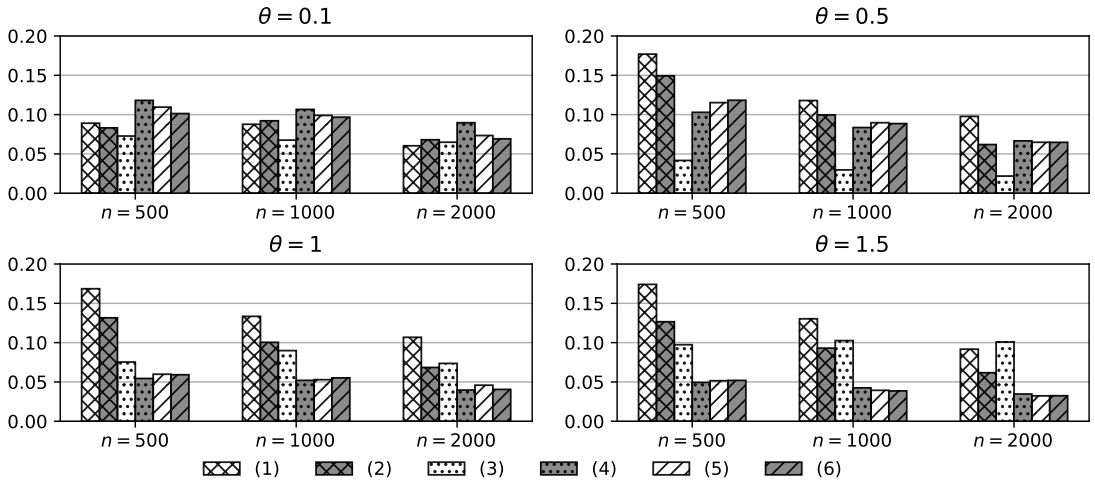
For the lowest theta parameter ($\theta = 1.1$) when the true TDC is 0, all the estimators exhibit similar results in terms of RMSE, which is also observed in the Gumbel simulations case. The

Figure 2.6: RMSE for the upper tail dependence with 100 Student simulations.



estimators (4), (5), and (6) also behave similarly and perform relatively well across all the parameterizations considered.

Figure 2.7: RMSE for the upper tail dependence with 100 rotated Clayton simulations.

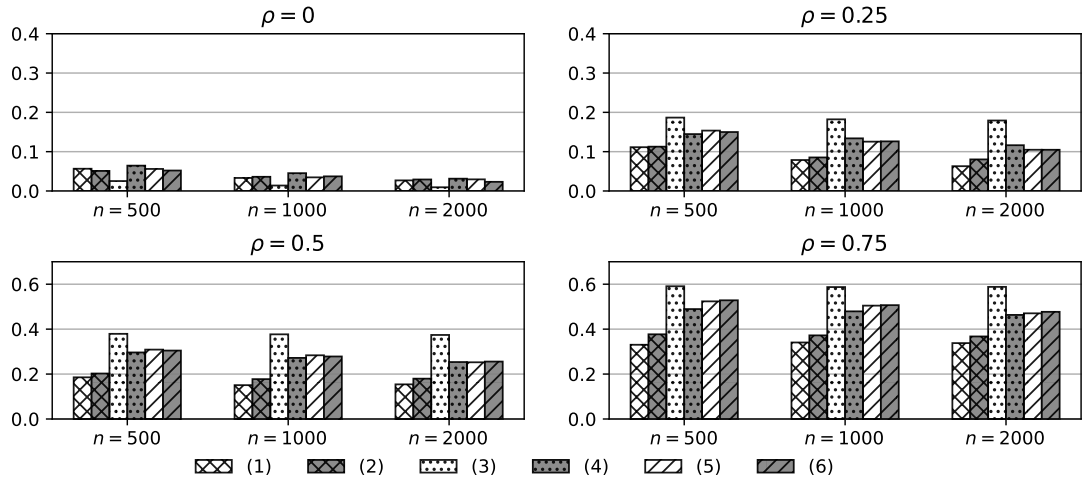


2.5.4 Gaussian simulations

We now evaluate the estimators on the Gaussian copula, that is in a framework with no tail dependence. We see that an increase in the correlation coefficient strongly biases all the estimators. Our results show that the nonparametric estimator for different threshold values captures tail dependence even when the true distribution does not exhibit tail dependence but only dependence for the bulk of the bivariate distribution. This is entirely consistent with Frahm et al. (2005b).

The results are gathered in Figure 2.8, with more details in Table 2.8 , in the appendix.

Figure 2.8: RMSE for the upper tail dependence with 100 Gaussian simulations.



2.5.5 Analysis

Overall, our two estimators based on a minimization of the theoretical MSE, that is estimators (5) and (6), are, with the plateau-finding algorithm (4), the best performing estimators. They are the least sensitive to a change of parameter in the generating distribution. In contrast, the performance of other estimators strongly depends on the distribution considered and on the sample size. For several of the examples above, we observe a slightly better performance for our introduced estimators with respect to the plateau-finding method, when n is large. This simulation study thus highlights the relevance of the estimators based on a minimization of the theoretical MSE.

However, when the true TDC is close to zero, most estimators, including the plateau-finding algorithm, perform poorly in terms of bias and RMSE, while their variance is relatively low. This result is consistent with the findings of Frahm et al. (2005b) and Poulin et al. (2007). Therefore, we suggest testing the tail dependence before computing an estimation of the TDC (Ledford and Tawn, 1996; Capéraà et al., 1997; Hoga, 2018).

2.6 Empirical application

Copulas are used in many fields, such as hydrology (Tawn, 1988; Genest and Favre, 2007; Poulin et al., 2007; Aghakouchak et al., 2010), astronomy (Scherrer et al., 2009; Sato et al., 2011), telecommunication networks (Garcin and Guégan, 2012; Neuhäuser et al., 2015). We focus here on the modelling of financial assets, whose literature also largely uses copulas. In particular, there is a prevalent use of TDCs to describe the dependence of extreme financial returns (Malevergne and Sornette, 2003; Poon et al., 2004; Caillault and Guégan, 2005).

In this short empirical application, we estimate the lower and upper TDCs using the six same estimators as used in the simulation study. We consider the MSCI developed markets indices,

which represent the performance of the overall financial securities in countries with a developed market. The estimation period is between 01/01/2000 and 01/01/2021. It consists of 5,295 daily observations for 18 countries. We estimate the pairwise TDCs between the US market and other developed markets. The results are reported in Tables 2.2 and 2.3, along with a standard linear dependence measure, namely the correlation coefficient.

Table 2.2: Lower TDC: US vs other developed countries.

Country	(1)	(2)	(3)	(4)	(5)	(6)	ρ
Canada	0.509	0.500	0.590	0.514	0.551	0.560	0.704
France	0.396	0.462	0.441	0.425	0.413	0.413	0.543
Germany	0.377	0.396	0.461	0.415	0.421	0.423	0.577
UK	0.415	0.415	0.422	0.378	0.387	0.389	0.528
Netherlands	0.340	0.415	0.431	0.401	0.373	0.380	0.535
Sweden	0.340	0.368	0.380	0.342	0.369	0.367	0.493
Belgium	0.340	0.434	0.344	0.344	0.350	0.351	0.470
Switzerland	0.377	0.368	0.326	0.366	0.355	0.351	0.458
Spain	0.377	0.358	0.368	0.353	0.336	0.336	0.494
Norway	0.396	0.396	0.278	0.322	0.358	0.365	0.427
Austria	0.396	0.377	0.257	0.329	0.341	0.339	0.401
Italy	0.340	0.321	0.370	0.310	0.327	0.324	0.499
Denmark	0.283	0.358	0.226	0.339	0.308	0.313	0.375
Australia	0.264	0.245	0.072	0.232	0.223	0.236	0.256
Singapore	0.245	0.217	0.120	0.196	0.219	0.209	0.281
Hong Kong	0.226	0.198	0.057	0.150	0.151	0.155	0.204
Japan	0.113	0.123	0.001	0.122	0.121	0.120	0.054

Notes: this table presents the lower TDC between the US and other developed countries, depending on the TDC estimator. The correlation coefficient is ρ.

We observe a global coherence between the results of the six estimators. The strongest discrepancies among the estimators appear for estimator (3), which tends to underestimate the lower TDC with respect to the other estimators, when the TDC is low, and which also tends to overestimate the upper TDC, except for the four countries with the lowest TDC. The arbitrary estimator (1) tends to underestimate the upper TDC.

The three Asian markets (Japan, Hong Kong, Singapore) exhibit the lowest lower and upper TDC. Of course, the time zone difference with the US can cause date shifts: if the US market is driving the global economy, the effect on the Asian markets is to be observed one day later. Even though the lower TDC is globally higher than the upper TDC, we see that a strong upper TDC is generally related to a strong lower TDC. This result suggests that one cannot benefit from a pairwise boom without having a risk of simultaneous crash.

2.7 Conclusion

We have given an expression of the MSE for the nonparametric TDC estimator. By minimizing this MSE in the case of a Clayton or a Gumbel copula, we have proposed a semiparametric method for estimating either the lower or the upper TDC. It is based on a plug-in approach, in which the pa-

Table 2.3: Upper TDC: US vs other developed countries.

Country	(1)	(2)	(3)	(4)	(5)	(6)	ρ
Canada	0.377	0.434	0.530	0.431	0.455	0.455	0.704
Germany	0.321	0.406	0.445	0.427	0.418	0.427	0.577
Netherlands	0.321	0.340	0.420	0.417	0.418	0.418	0.535
France	0.283	0.340	0.426	0.317	0.404	0.410	0.543
UK	0.302	0.321	0.405	0.313	0.332	0.353	0.528
Sweden	0.302	0.330	0.382	0.291	0.335	0.330	0.493
Spain	0.283	0.292	0.372	0.295	0.320	0.313	0.494
Belgium	0.264	0.302	0.356	0.284	0.330	0.330	0.470
Italy	0.264	0.264	0.384	0.256	0.300	0.302	0.458
Switzerland	0.302	0.236	0.336	0.242	0.269	0.268	0.499
Norway	0.226	0.274	0.302	0.274	0.264	0.268	0.427
Denmark	0.226	0.236	0.267	0.214	0.251	0.253	0.375
Austria	0.132	0.236	0.282	0.240	0.257	0.257	0.401
Singapore	0.226	0.264	0.200	0.219	0.233	0.233	0.281
Australia	0.189	0.245	0.164	0.227	0.216	0.216	0.256
Hong Kong	0.151	0.151	0.158	0.158	0.165	0.163	0.204
Japan	0.075	0.113	0.052	0.103	0.105	0.105	0.054

Notes: this table presents the upper TDC between the US and other developed countries, depending on the TDC estimator. The correlation coefficient is ρ.

rameter of the Clayton or Gumbel copula is estimated on a well-chosen part of the observations. A simulation study shows that this kind of estimator offers better performance. It behaves in a similar way as the method relying on the plateau-finding algorithm does, for data generated by various types of copulas. These results, along with the relative simplicity of the method compared to the plateau-finding algorithm, thus legitimize this new estimator. Therefore we recommend our estimators when facing an unknown type of underlying distribution.

2.A Appendix

2.A.1 Proof of Theorem 2.3.1

We define a dependence parameter for a given probability q as:

$$\lambda_L(q) = \frac{\delta(q)}{q}. \tag{A.1}$$

In particular, $\lambda_L = \lim_{q\to 0} \lambda_L(q)$. Since $\delta(0) = 0$ by a basic property of copulas, we also have $\lambda_L = \delta'(0)$. We decompose the error of the estimator provided in equation (1) in noise and bias:

$$\widehat{\lambda}_{L,n}\left(\frac{i(n)}{n}\right) - \lambda_L = \left(\widehat{\lambda}_{L,n}\left(\frac{i(n)}{n}\right) - \lambda_L\left(\frac{i(n)}{n}\right)\right) + \left(\lambda_L\left(\frac{i(n)}{n}\right) - \lambda_L\right).$$

Following the work of Fermanian *et al.* and in particular their Theorem 3, we know that the empirical copula process $\sqrt{n}(\widehat{C}_n(u,v) - C(u,v))$ converges weakly towards a Gaussian process

$G_C(u,v)$ (Fermanian et al., 2004):

$$G_C(u,v) = B_C(u,v) - h_1(u,v)B_C(u,1) - h_2(u,v)B_C(1,v),$$

where B_C is a Brownian bridge on $[0,1]^2$ of covariance

$$\left[B_C(u,v)B_C(u',v')\right] = C(u \wedge u', v \wedge v') - C(u,v)C(u',v'). \tag{A.2}$$

Therefore, $\sqrt{n}\left(\widehat{\lambda}_{L,n}(i(n)/n) - \lambda_L(i(n)/n)\right) = \sqrt{n}(\widehat{C}_n(\alpha,\alpha) - C(\alpha,\alpha))/\alpha$, weakly converges toward $G_C(\alpha,\alpha)/\alpha$. As a consequence, $n\left(\widehat{\lambda}_{L,n}(i(n)/n) - \lambda_L(i(n)/n)\right)^2$ weakly converges toward $G_C(\alpha,\alpha)^2/\alpha^2$ (**?**, Th. 5.2) and, thanks to the assumed uniform integrability, $\left[n\left(\widehat{\lambda}_{L,n}(i(n)/n) - \lambda_L(i(n)/n)\right)^2\right]$ converges toward $\left[G_C(\alpha,\alpha)^2\right]/\alpha^2$ (**?**, Th. 5.4). We note that, for $u \in (0,1)$,

$$\begin{aligned}
[G_C(u,u)^2] &= [B_C(u,u)^2] + h_1(u)^2[B_C(u,1)^2] + h_2(u)^2[B_C(1,u)^2] \\
&\quad -2h_1(u)[B_C(u,u)B_C(u,1)] - 2h_2(u)[B_C(u,u)B_C(1,u)] \\
&\quad +2h_1(u)h_2(u)[B_C(u,1)B_C(1,u)] \\
&= \delta(u) - \delta(u)^2 + h_1(u)^2\left(C(u,1) - C(u,1)^2\right) + h_2(u)^2\left(C(1,u) - C(1,u)^2\right) \\
&\quad -2h_1(u)\left(\delta(u) - \delta(u)C(u,1)\right) - 2h_2(u)\left(\delta(u) - \delta(u)C(1,u)\right) \\
&\quad +2h_1(u)h_2(u)\left(\delta(u) - C(u,1)C(1,u)\right) \\
&= \delta(u)(1-\delta(u)) + h_1(u)^2u(1-u) + h_2(u)^2u(1-u) \\
&\quad -2h_1(u)\delta(u)(1-u) - 2h_2(u)\delta(u)(1-u) \\
&\quad +2h_1(u)h_2(u)\left(\delta(u) - u^2\right) \\
&= \sigma^2(u),
\end{aligned}$$

according to equation (A.2) and using the fact that $C(u,1) = C(1,u) = u$. As a consequence, the asymptotic MSE is

$$\frac{1}{n\alpha^2}\left[G_C(\alpha,\alpha)^2\right] + (\lambda_L(\alpha) - \lambda_L)^2 = \frac{1}{n\alpha^2}\sigma^2(\alpha) + \left(\frac{1}{\alpha}\delta(\alpha) - \delta'(0)\right)^2.$$

2.A.2 Proof of Proposition 2.3.1

The diagonal section of the Clayton copula is obtained by considering the case $u = v$:

$$\delta(u) = \left(2u^{-\theta} - 1\right)^{-1/\theta}.$$

Its first derivative is:

$$\delta'(u) = 2u^{-\theta-1}\left(2u^{-\theta} - 1\right)^{-1-1/\theta} = 2\left(2 - u^{\theta}\right)^{-1-1/\theta},$$

whose value in $u = 0$ is $\delta'(0) = 2^{-1/\theta}$. According to Theorem 2.3.1, the asymptotic bias is thus:

$$\frac{1}{\alpha}\delta(\alpha) - \delta'(0) = \frac{1}{\alpha}\left(2\alpha^{-\theta} - 1\right)^{-1/\theta} - 2^{-1/\theta} = \left(2 - \alpha^{\theta}\right)^{-1/\theta} - 2^{-1/\theta}.$$

The corresponding h-function is (Schepsmeier and Stöber, 2014):

$$h(u) = \frac{\delta'(u)}{2} = \delta(u)\frac{u^{-\theta-1}}{(2u^{-\theta}-1)} = \frac{\delta(u)}{u}\frac{1}{2-u^{\theta}}.$$

Following equation (3), we get:

$$\begin{aligned}
\sigma^2(u) &= \delta(u)(1-\delta(u)) + 2(1-u)h(u)\left[uh(u) - 2\delta(u)\right] + 2h(u)^2(\delta(u)-u^2) \\
&= \delta(u)(1-\delta(u)) + 2(1-u)\tfrac{\delta(u)}{u}\tfrac{1}{2-u^{\theta}}\left[u\tfrac{\delta(u)}{u}\tfrac{1}{2-u^{\theta}} - 2\delta(u)\right] + 2\left(\tfrac{\delta(u)}{u}\tfrac{1}{2-u^{\theta}}\right)^2(\delta(u)-u^2) \\
&= \delta(u)(1-\delta(u)) + 2\left(\tfrac{1}{u}-1\right)\delta(u)^2\tfrac{1}{2-u^{\theta}}\left[\tfrac{1}{2-u^{\theta}} - 2\right] + 2\tfrac{\delta(u)^2}{(2-u^{\theta})^2}\left(\tfrac{\delta(u)}{u^2}-1\right) \\
&= \delta(u) + \delta(u)^2\left[-1 + 2\left(\tfrac{1}{u}-1\right)\tfrac{1}{2-u^{\theta}}\left(\tfrac{1}{2-u^{\theta}} - 2\right) - \tfrac{2}{(2-u^{\theta})^2}\right] + 2\delta(u)^3\left[\tfrac{1}{u^2(2-u^{\theta})^2}\right] \\
&= \delta(u) - \delta(u)^2\left[1 + 2\tfrac{2(1-u)(1-u^{\theta})+1}{u(2-u^{\theta})^2}\right] + 2\delta(u)^3\left[\tfrac{1}{u^2(2-u^{\theta})^2}\right].
\end{aligned}$$

2.A.3 Proof of Theorem 2.3.2

We first focus on the variance part of the tail dependence estimator:

$$\left[\left(\widehat{\lambda}_{U,n}(i(n)/n) - \lambda_U(i(n)/n)\right)^2\right].$$

We know that $\sqrt{n}(\widehat{C}(u,v) - C(u,v))$ converges weakly towards a Gaussian process $G_C(u,v)$, as already mentioned in the proof of Theorem 2.3.1. Therefore, following the same reasoning as in the proof of Theorem 2.3.1, the second moment of $\widehat{\lambda}_U(i(n)/n) - \lambda_U(i(n)/n)$ converges toward the second moment of $G_C(\alpha,\alpha)\sqrt{n}/(n-\alpha n)$, which is of mean 0 and of variance

$$\frac{1}{1(1-\alpha)^2}\mathbb{E}\left[G_C(\alpha,\alpha)^2\right] = \frac{1}{n(1-\alpha)^2}\sigma^2(\alpha).$$

By noting that $\lambda_U = 2 - \delta'(1)$, the expression of the bias is straightforward and we conclude by noting that the MSE is the sum of the variance and of the squared bias.

2.A.4 Proof of Proposition 2.3.2

In the Gumbel case,

$$\delta(u) = C(u,u) = \exp\left[-\left\{2\left(-\ln(u)\right)^{\theta}\right\}^{\frac{1}{\theta}}\right] = \exp\left[-(2t)^{\frac{1}{\theta}}\right],$$

where $t = (-\ln(u))^{\theta}$. Its first derivative is $\delta'(u) = \frac{2^{1/\theta}\delta(u)}{u}$, and in particular $\delta'(1) = 2^{1/\theta}$. According to Theorem 2.3.2, the asymptotic bias is thus:

$$\frac{1-2\alpha+\delta(\alpha)}{1-\alpha} - 2 + \delta'(1) = \frac{1-2\alpha+\delta(\alpha)}{1-\alpha} - 2 + 2^{1/\theta}.$$

The h-function is (Schepsmeier and Stöber, 2014):

$$h(u) = -\frac{\mathrm{e}^{-(2t)^{\frac{1}{\theta}}}(2t)^{\frac{1}{\theta}-1}t}{u\ln(u)} = \frac{\delta(u)}{u}2^{\frac{1}{\theta}-1},$$

so that

$$\begin{aligned}\sigma^2(u) &= \delta(u)(1-\delta(u)) + 2(1-u)h(u)[uh(u) - 2\delta(u)] + 2h(u)^2\left(\delta(u) - u^2\right) \\ &= \delta(u)(1-\delta(u)) + \delta(u)^2\left(\tfrac{1}{u} - 1\right)2^{\frac{1}{\theta}}\left(2^{\frac{1}{\theta}-1} - 2\right) + \delta(u)^2 2^{\frac{2}{\theta}-1}\left(\tfrac{\delta(u)}{u^2} - 1\right).\end{aligned}$$

2.A.5 Proof of Theorem 2.3.3

Like in Theorems 2.3.1 and 2.3.2, we decompose the MSE in variance and squared bias:

$$\begin{aligned}\left[\left(\tfrac{1}{m}\textstyle\sum_{k=1}^m \widehat{\lambda}_L\left(\tfrac{i_k(n)}{n}\right) - \lambda_L\right)^2\right] &= \left[\tfrac{1}{m}\textstyle\sum_{k,l=1}^m \left(\widehat{\lambda}_L\left(\tfrac{i_k(n)}{n}\right) - \lambda_L\left(\tfrac{i_k(n)}{n}\right)\right)\left(\widehat{\lambda}_L\left(\tfrac{i_l(n)}{n}\right) - \lambda_L\left(\tfrac{i_l(n)}{n}\right)\right)\right] \\ &\quad + \left[\left(\tfrac{1}{m}\textstyle\sum_{k=1}^m \lambda_L(\alpha_k) - \lambda_L\right)^2\right],\end{aligned}$$

with $\lambda_L(u)$ defined by equation (A.1). Like in the previous theorems, the bias part finds a straightforward expression using δ and δ':

$$\left(\frac{1}{m}\sum_{k=1}^m \lambda_L(\alpha_k) - \lambda_L\right)^2 = \left(\frac{1}{m}\sum_{k=1}^m \frac{1}{\alpha_k}\delta(\alpha_k) - \delta'(0)\right)^2.$$

We now focus on the variance part of the MSE. Like in the proof of Theorem 2.3.1, we note that the empirical copula process $\sqrt{n}(\widehat{C}(u,v) - C(u,v))$ converges weakly towards a Gaussian process $G_C(u,v)$:

$$G_C(u,v) = B_C(u,v) - h_1(u,v)B_C(u,1) - h_2(u,v)B_C(1,v),$$

where B_C is a Brownian bridge on $[0,1]^2$ whose covariance is provided by equation (A.2). Therefore, the covariance between $G_C(u,u)$ and $G_C(v,v)$ is

$$\begin{aligned}[G_C(u,u)G_C(v,v)] &= [B_C(u,u)B_C(v,v)] + h_1(u)h_1(v)[B_C(u,1)B_C(v,1)] + h_2(u)h_2(v)[B_C(1,u)B_C(1,v)] \\ &\quad - h_1(v)[B_C(u,u)B_C(v,1)] - h_2(v)[B_C(u,u)B_C(1,v)] \\ &\quad - h_1(u)[B_C(v,v)B_C(u,1)] - h_2(u)[B_C(v,v)B_C(1,u)] \\ &\quad + h_1(u)h_2(v)[B_C(u,1)B_C(1,v)] + h_1(v)h_2(u)[B_C(v,1)B_C(1,u)] \\ &= \delta(u\wedge v) - \delta(u)\delta(v) + (h_1(u)h_1(v) + h_2(u)h_2(v))((u\wedge v) - uv) \\ &\quad - h_1(v)(C(u\wedge v,u) - v\delta(u)) - h_2(v)(C(u,u\wedge v) - v\delta(u)) \\ &\quad - h_1(u)(C(u\wedge v,v) - u\delta(v)) - h_2(u)(C(v,u\wedge v) - u\delta(v)) \\ &\quad + h_1(u)h_2(v)(C(u,v) - uv) + h_1(v)h_2(u)(C(v,u) - uv)) \\ &= \mathcal{K}(u,v),\end{aligned}$$

according to equation (A.2) and using the fact that $C(u,1) = C(1,u) = u$. Following the same reasoning as in the proof of Theorem 2.3.1, this result leads to the expression of the asymptotic variance of the TDC estimator provided in Theorem 2.3.3 thanks to equation (A.1), which links $(\widehat{\lambda}_L(u) - \lambda_L(u))$ to $(\widehat{\delta}(u) - \delta(u))/u$, that is to $(\widehat{C}(u,u) - C(u,u))/u$.

2.A.6 Uniform integrability condition of Theorem 2.3.1 for the independent copula

In Theorem 2.3.1, we have assumed that $n\left(\widehat{\delta}_n(\alpha) - \delta(\alpha)\right)^2$ is uniformly integrable. We now show that this assumption is fulfilled in the case of the independent copula. It is in fact enough to show

that (Billingsley, 2013, page 32):

$$\exists \varepsilon > 0,\ \sup_n \left(\left| n\left(\widehat{\delta}_n(\alpha) - \delta(\alpha)\right)^2 \right|^{1+\varepsilon} \right) < \infty. \tag{A.3}$$

For simplicity of the notations, and without any consequence on the final result, we assume that αn is an integer. For the independent copula, we have (Deheuvels, 1979, Th. 4.1):

$$\left[\widehat{\delta}_n(\alpha)\right] = \delta(\alpha) = \alpha^2 \tag{A.4}$$

and

$$\left[\widehat{\delta}_n(\alpha)^2\right] = \frac{\alpha^2}{n} + \frac{\alpha^2(\alpha n - 1)^2}{n(n-1)}. \tag{A.5}$$

Regarding the third moment, we follow a similar proof as the one proposed by (Deheuvels, 1979, Th. 4.1). We first define the rank statistics $r_{i,X}$ and $r_{i,Y}$, such that $X_{r_{1,X}} \leq X_{r_{2,X}} \leq ... \leq X_{r_{n,X}}$ and $Y_{r_{1,Y}} \leq Y_{r_{2,Y}} \leq ... \leq X_{r_{n,Y}}$. Then,

$$\widehat{\delta}_n(\alpha) = \frac{1}{n} \sum_{i=1}^{N} \left\{ r_{i,X} \leq \alpha n \right\} \left\{ r_{i,Y} \leq \alpha n \right\}.$$

Therefore, by independence between the two components X and Y, and noting

$$p_{i,j,k}(z) = \left(r_{i,X} \leq z, r_{j,X} \leq z, r_{k,X} \leq z \right),$$

we obtain

$$\begin{array}{rcl} \left[\widehat{\delta}_n(\alpha)^3\right] & = & \frac{1}{n^3} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} p_{i,j,k}(\alpha n)^2 \\ & = & \frac{1}{n^3} \left(\sum_{i=j=k} p_{i,j,k}(\alpha n)^2 + 3 \sum_{i=j\neq k} p_{i,j,k}(\alpha n)^2 + \sum_{i\neq j\neq k} p_{i,j,k}(\alpha n)^2 \right). \end{array} \tag{A.6}$$

Moreover, for $m \in [\![1, n]\!]$,

$$p_{i,i,i}(m) = \sum_{m'=1}^{m} \left(r_{i,X} = m' \right) = \sum_{m'=1}^{m} \frac{1}{n} = \frac{m}{n}.$$

For $m < m'$, we have $\sum_{i\neq j} (r_{i,X} = m, r_{j,X} = m') = 1$, with all the events of the form $(r_{i,X} = m, r_{j,X} = m')$ having equal probability, so that $(r_{i,X} = m, r_{j,X} = m') = 1/n(n-1)$. Similarly, for $m < m' < m''$, $(r_{i,X} = m, r_{j,X} = m', r_{k,X} = m'') = 1/n(n-1)(n-2)$. As a consequence, we have

$$p_{i,i,j}(m) = \sum_{m'=1}^{m} \sum_{m''=1}^{m} \left(r_{i,X} = m', r_{j,X} = m'' \right) = \frac{m(m-1)}{n(n-1)}$$

for $i \neq j$ (and therefore $m' \neq m''$). Similarly, for $i \neq j \neq k$, we have

$$p_{i,j,k}(m) = \frac{m(m-1)(m-2)}{n(n-1)(n-2)}.$$

Since $p_{i,i,i}(m)$, $p_{i,i,j}(m)$, and $p_{i,j,k}(m)$ do not depend on the exact value of i, j, and k, equation (A.6) becomes

$$\begin{aligned}\left[\widehat{\delta}_n(\alpha)^3\right] &= \tfrac{1}{n^3}\left(np_{1,1,1}(\alpha n)^2+3n(n-1)p_{1,1,2}(\alpha n)^2+n(n-1)(n-2)p_{1,2,3}(\alpha n)^2\right)\\ &= \tfrac{1}{n^3}\left(n\left[\tfrac{\alpha n}{n}\right]^2+3n(n-1)\left[\tfrac{\alpha n(\alpha n-1)}{n(n-1)}\right]^2+n(n-1)(n-2)\left[\tfrac{\alpha n(\alpha n-1)(\alpha n-2)}{n(n-1)(n-2)}\right]^2\right)\\ &= \tfrac{\alpha^2}{n^2}+3\tfrac{\alpha^2(\alpha n-1)^2}{n^2(n-1)}+\tfrac{\alpha^2(\alpha n-1)^2(\alpha n-2)^2}{n^2(n-1)(n-2)}.\end{aligned} \tag{A.7}$$

In a very similar way, after introducing $p_{i,j,k,l}(z) = \left(r_{i,X} \le z, r_{j,X} \le z, r_{k,X} \le z, r_{l,X} \le z\right)$, we can prove that

$$\begin{aligned}\left[\widehat{\delta}_n(\alpha)^4\right] &= \tfrac{1}{n^4}\left(\textstyle\sum_{i=j=k=l} p_{i,j,k,l}(\alpha n)^2+4\sum_{i=j=k\neq l} p_{i,j,k,l}(\alpha n)^2+3\sum_{i=j\neq k=l} p_{i,j,k,l}(\alpha n)^2\right.\\ &\quad \left.+6\textstyle\sum_{i=j\neq k\neq l} p_{i,j,k,l}(\alpha n)^2+\sum_{i\neq j\neq k\neq l} p_{i,j,k,l}(\alpha n)^2\right)\\ &= \tfrac{\alpha^2}{n^3}+7\tfrac{\alpha^2(\alpha n-1)^2}{n^3(n-1)}+6\tfrac{\alpha^2(\alpha n-1)^2(\alpha n-2)^2}{n^3(n-1)(n-2)}+\tfrac{\alpha^2(\alpha n-1)^2(\alpha n-2)^2(\alpha n-3)^2}{n^3(n-1)(n-2)(n-3)}.\end{aligned} \tag{A.8}$$

Finally, using equations (A.4), (A.5), (A.7), and (A.8), and simplifying some long expression, we obtain:

$$\begin{aligned}\left[n^2\left(\widehat{\delta}_n(\alpha)-\delta(\alpha)\right)^4\right] &= n^2\left(\left[\widehat{\delta}_n(\alpha)^4\right]-4\left[\widehat{\delta}_n(\alpha)^3\right]\delta(\alpha)+6\left[\widehat{\delta}_n(\alpha)^2\right]\delta(\alpha)^2-4\left[\widehat{\delta}_n(\alpha)\right]\delta(\alpha)^3+\delta(\alpha)^4\right)\\ &= \tfrac{A_3(\alpha)n^3+A_2(\alpha)n^2+A_1(\alpha)n}{(n-1)(n-2)(n-3)},\end{aligned}$$

with

$$\begin{cases} A_3(\alpha) &= 3\alpha^8-12\alpha^7+18\alpha^6-12\alpha^5+3\alpha^4\\ A_2(\alpha) &= 18\alpha^8-72\alpha^7+120\alpha^6-108\alpha^5+55\alpha^4-14\alpha^3+\alpha^2\\ A_1(\alpha) &= \alpha^4-2\alpha^3+\alpha^2,\end{cases}$$

so that $\left[n^2\left(\widehat{\delta}_n(\alpha)-\delta(\alpha)\right)^4\right] < \infty$ whatever n. Using the condition put forward in equation (A.3), this leads to the uniform integrability of $n\left(\widehat{\delta}_n(\alpha)-\delta(\alpha)\right)^2$.

2.A.7 Tables of results for the simulation study

Table 2.4 : Upper tail dependence with 100 Gumbel simulations.

Dataset	Method	$n = 500$			$n = 1000$			$n = 2000$		
		Bias	$\sigma(\widehat{\lambda}_{U,n})$	RMSE	Bias	$\sigma(\widehat{\lambda}_{U,n})$	RMSE	Bias	$\sigma(\widehat{\lambda}_{U,n})$	RMSE
$\theta = 1.10$	(1)	0.08	0.19	0.21	0.09	0.11	0.14	0.09	0.09	0.12
$\lambda_U = 0.12$	(2)	0.11	0.12	0.17	0.09	0.08	0.12	0.12	0.06	0.13
	(3)	0.15	0.03	0.15	0.15	0.03	0.15	0.15	0.02	0.15
	(4)	0.18	0.06	0.19	0.17	0.06	0.18	0.16	0.06	0.17
	(5)	0.18	0.07	0.20	0.17	0.06	0.18	0.16	0.04	0.16
	(6)	0.18	0.07	0.20	0.17	0.06	0.18	0.15	0.04	0.16
$\theta = 1.50$	(1)	-0.24	0.17	0.29	-0.22	0.12	0.25	-0.22	0.08	0.23
$\lambda_U = 0.41$	(2)	-0.19	0.12	0.23	-0.20	0.08	0.22	-0.19	0.07	0.20
	(3)	-0.10	0.03	0.10	-0.09	0.02	0.10	-0.09	0.02	0.10
	(4)	-0.12	0.06	0.13	-0.13	0.06	0.14	-0.14	0.05	0.15
	(5)	-0.11	0.07	0.13	-0.13	0.06	0.14	-0.14	0.04	0.15
	(6)	-0.11	0.06	0.13	-0.13	0.06	0.14	-0.14	0.04	0.15
$\theta = 1.75$	(1)	-0.32	0.17	0.36	-0.29	0.13	0.31	-0.29	0.08	0.30
$\lambda_U = 0.51$	(2)	-0.25	0.14	0.29	-0.25	0.09	0.26	-0.26	0.06	0.26
	(3)	-0.12	0.03	0.12	-0.12	0.02	0.12	-0.12	0.02	0.12
	(4)	-0.16	0.08	0.18	-0.18	0.07	0.19	-0.18	0.06	0.19
	(5)	-0.14	0.07	0.16	-0.16	0.05	0.17	-0.18	0.04	0.19
	(6)	-0.14	0.06	0.16	-0.16	0.05	0.17	-0.18	0.04	0.18
$\theta = 2.00$	(1)	-0.28	0.20	0.34	-0.28	0.12	0.30	-0.28	0.08	0.30
$\lambda_U = 0.59$	(2)	-0.24	0.14	0.28	-0.25	0.10	0.27	-0.25	0.07	0.26
	(3)	-0.11	0.02	0.11	-0.11	0.02	0.11	-0.11	0.01	0.11
	(4)	-0.14	0.06	0.16	-0.17	0.06	0.18	-0.18	0.06	0.19
	(5)	-0.13	0.07	0.15	-0.15	0.05	0.16	-0.17	0.04	0.17
	(6)	-0.12	0.06	0.14	-0.15	0.05	0.16	-0.17	0.04	0.17

Table 2.5 : Upper tail dependence with 100 Student simulations, with $\rho = 0$.

Dataset	Method	$n = 500$			$n = 1000$			$n = 2000$		
		Bias	$\sigma(\widehat{\lambda}_{U,n})$	RMSE	Bias	$\sigma(\widehat{\lambda}_{U,n})$	RMSE	Bias	$\sigma(\widehat{\lambda}_{U,n})$	RMSE
$\rho = 0$	(1)	0.00	0.19	0.19	-0.01	0.13	0.13	-0.01	0.10	0.10
$\nu = 1$	(2)	-0.01	0.13	0.13	0.01	0.08	0.09	-0.01	0.07	0.07
$\lambda_U = 0.29$	(3)	-0.19	0.04	0.19	-0.19	0.03	0.19	-0.19	0.02	0.19
	(4)	-0.02	0.07	0.07	-0.00	0.05	0.05	-0.01	0.05	0.05
	(5)	-0.01	0.09	0.09	0.01	0.06	0.06	-0.00	0.06	0.06
	(6)	-0.01	0.08	0.09	0.01	0.06	0.06	-0.00	0.06	0.06
$\rho = 0$	(1)	-0.04	0.14	0.14	0.01	0.11	0.11	-0.02	0.08	0.08
$\nu = 2$	(2)	-0.02	0.10	0.10	0.00	0.08	0.08	-0.00	0.06	0.06
$\lambda_U = 0.18$	(3)	-0.11	0.04	0.12	-0.11	0.02	0.12	-0.11	0.02	0.11
	(4)	0.00	0.07	0.07	0.00	0.05	0.05	0.01	0.05	0.05
	(5)	-0.00	0.09	0.09	0.00	0.07	0.07	0.01	0.06	0.06
	(6)	-0.00	0.09	0.09	0.00	0.07	0.07	0.01	0.06	0.06
$\rho = 0$	(1)	-0.01	0.13	0.13	0.01	0.11	0.11	0.00	0.07	0.07
$\nu = 3$	(2)	0.00	0.10	0.10	0.01	0.07	0.07	0.01	0.05	0.05
$\lambda_U = 0.12$	(3)	-0.06	0.04	0.07	-0.06	0.02	0.06	-0.06	0.02	0.07
	(4)	0.03	0.07	0.08	0.04	0.05	0.06	0.02	0.04	0.04
	(5)	0.01	0.08	0.09	0.02	0.06	0.07	0.02	0.05	0.06
	(6)	0.01	0.09	0.09	0.02	0.06	0.07	0.02	0.05	0.05

Table 2.6 : Upper tail dependence with 100 Student simulations, with $\rho = 0.25$.

Dataset	Method	$n = 500$			$n = 1000$			$n = 2000$		
		Bias	$\sigma(\widehat{\lambda}_{U,n})$	RMSE	Bias	$\sigma(\widehat{\lambda}_{U,n})$	RMSE	Bias	$\sigma(\widehat{\lambda}_{U,n})$	RMSE
$\rho = 0.25$	(1)	-0.01	0.20	0.20	-0.01	0.13	0.13	-0.02	0.10	0.11
$\nu = 1$	(2)	-0.02	0.14	0.14	0.01	0.09	0.09	-0.01	0.07	0.07
$\lambda_U = 0.39$	(3)	-0.12	0.04	0.13	-0.12	0.03	0.13	-0.12	0.02	0.12
	(4)	-0.02	0.07	0.07	-0.01	0.05	0.05	-0.01	0.05	0.05
	(5)	-0.00	0.07	0.07	0.00	0.05	0.05	-0.00	0.04	0.04
	(6)	-0.00	0.07	0.07	0.00	0.05	0.05	-0.00	0.04	0.04
$\rho = 0.25$	(1)	-0.03	0.16	0.17	-0.00	0.12	0.12	-0.00	0.09	0.09
$\nu = 2$	(2)	-0.03	0.11	0.11	-0.00	0.10	0.10	0.01	0.07	0.07
$\lambda_U = 0.27$	(3)	-0.04	0.04	0.06	-0.04	0.03	0.05	-0.04	0.02	0.04
	(4)	0.01	0.06	0.06	0.01	0.05	0.05	0.01	0.05	0.05
	(5)	0.02	0.08	0.08	0.01	0.05	0.05	0.02	0.05	0.05
	(6)	0.02	0.07	0.08	0.01	0.05	0.05	0.02	0.05	0.05
$\rho = 0.25$	(1)	-0.00	0.16	0.16	0.01	0.13	0.13	-0.00	0.08	0.08
$\nu = 3$	(2)	0.00	0.12	0.12	0.03	0.09	0.10	0.02	0.06	0.06
$\lambda_U = 0.20$	(3)	0.02	0.04	0.05	0.03	0.03	0.04	0.02	0.02	0.03
	(4)	0.04	0.08	0.09	0.05	0.06	0.07	0.04	0.05	0.06
	(5)	0.06	0.09	0.10	0.06	0.05	0.08	0.04	0.04	0.06
	(6)	0.05	0.08	0.10	0.06	0.05	0.07	0.04	0.04	0.06

Table 2.7 : Upper tail dependence with 100 rotated Clayton simulations.

Dataset	Method	$n = 500$			$n = 1000$			$n = 2000$		
		Bias	$\sigma(\widehat{\lambda}_{U,n})$	RMSE	Bias	$\sigma(\widehat{\lambda}_{U,n})$	RMSE	Bias	$\sigma(\widehat{\lambda}_{U,n})$	RMSE
$\theta = 0.1$	(1)	0.04	0.08	0.09	0.06	0.07	0.09	0.04	0.04	0.06
$\lambda_U = 0.00$	(2)	0.05	0.07	0.08	0.07	0.06	0.09	0.06	0.04	0.07
	(3)	0.06	0.04	0.07	0.06	0.02	0.07	0.06	0.02	0.06
	(4)	0.10	0.07	0.12	0.09	0.05	0.11	0.08	0.04	0.09
	(5)	0.08	0.07	0.11	0.08	0.05	0.10	0.07	0.03	0.07
	(6)	0.08	0.07	0.10	0.08	0.05	0.10	0.06	0.03	0.07
$\theta = 0.5$	(1)	0.01	0.18	0.18	0.01	0.12	0.12	0.02	0.10	0.10
$\lambda_U = 0.25$	(2)	0.04	0.14	0.15	0.04	0.09	0.10	0.03	0.05	0.06
	(3)	0.02	0.04	0.04	0.02	0.02	0.03	0.02	0.02	0.02
	(4)	0.08	0.07	0.10	0.06	0.06	0.08	0.05	0.05	0.07
	(5)	0.08	0.08	0.12	0.07	0.05	0.09	0.05	0.04	0.06
	(6)	0.09	0.08	0.12	0.07	0.05	0.09	0.05	0.04	0.06
$\theta = 1.0$	(1)	-0.00	0.17	0.17	-0.02	0.13	0.13	-0.02	0.10	0.11
$\lambda_U = 0.50$	(2)	-0.00	0.13	0.13	0.01	0.10	0.10	-0.01	0.07	0.07
	(3)	-0.07	0.03	0.08	-0.08	0.05	0.09	-0.07	0.01	0.07
	(4)	0.00	0.05	0.05	0.00	0.05	0.05	-0.00	0.04	0.04
	(5)	0.03	0.05	0.06	0.02	0.05	0.05	0.01	0.04	0.05
	(6)	0.02	0.06	0.06	0.02	0.05	0.06	0.01	0.04	0.04
$\theta = 1.5$	(1)	-0.02	0.17	0.17	-0.01	0.13	0.13	0.00	0.09	0.09
$\lambda_U = 0.63$	(2)	-0.04	0.12	0.13	-0.02	0.09	0.09	-0.00	0.06	0.06
	(3)	-0.10	0.02	0.10	-0.10	0.02	0.10	-0.10	0.01	0.10
	(4)	-0.02	0.04	0.05	-0.02	0.04	0.04	-0.01	0.03	0.03
	(5)	0.01	0.05	0.05	0.00	0.04	0.04	0.00	0.03	0.03
	(6)	0.01	0.05	0.05	0.00	0.04	0.04	0.00	0.03	0.03

Table 2.8 : Upper tail dependence with 100 Gaussian simulations.

Dataset	Method	$n = 500$			$n = 1000$			$n = 2000$		
		Bias	$\sigma(\widehat{\lambda}_{U,n})$	RMSE	Bias	$\sigma(\widehat{\lambda}_{U,n})$	RMSE	Bias	$\sigma(\widehat{\lambda}_{U,n})$	RMSE
$\rho = 0.00$	(1)	0.02	0.05	0.06	0.01	0.03	0.03	0.01	0.02	0.03
	(2)	0.02	0.04	0.05	0.02	0.03	0.04	0.02	0.02	0.03
	(3)	0.01	0.02	0.03	0.01	0.01	0.01	0.01	0.01	0.01
	(4)	0.04	0.05	0.06	0.03	0.04	0.05	0.02	0.02	0.03
	(5)	0.03	0.05	0.06	0.02	0.03	0.03	0.01	0.03	0.03
	(6)	0.02	0.05	0.05	0.02	0.03	0.04	0.01	0.02	0.02
$\rho = 0.25$	(1)	0.05	0.10	0.11	0.05	0.06	0.08	0.05	0.04	0.06
	(2)	0.08	0.08	0.11	0.07	0.05	0.09	0.07	0.04	0.08
	(3)	0.18	0.04	0.19	0.18	0.02	0.18	0.18	0.02	0.18
	(4)	0.12	0.07	0.14	0.12	0.06	0.13	0.11	0.05	0.12
	(5)	0.14	0.07	0.15	0.12	0.05	0.13	0.10	0.03	0.10
	(6)	0.14	0.07	0.15	0.12	0.05	0.13	0.10	0.03	0.10
$\rho = 0.50$	(1)	0.12	0.14	0.19	0.12	0.09	0.15	0.14	0.07	0.15
	(2)	0.17	0.11	0.20	0.16	0.07	0.18	0.17	0.06	0.18
	(3)	0.38	0.03	0.38	0.38	0.02	0.38	0.37	0.02	0.37
	(4)	0.28	0.08	0.30	0.26	0.07	0.27	0.24	0.07	0.25
	(5)	0.30	0.07	0.31	0.28	0.06	0.28	0.25	0.04	0.25
	(6)	0.30	0.06	0.30	0.27	0.05	0.28	0.25	0.04	0.26
$\rho = 0.75$	(1)	0.29	0.16	0.33	0.32	0.13	0.34	0.32	0.10	0.34
	(2)	0.35	0.13	0.38	0.36	0.09	0.37	0.36	0.07	0.37
	(3)	0.59	0.02	0.59	0.58	0.06	0.59	0.59	0.01	0.59
	(4)	0.48	0.06	0.49	0.47	0.08	0.48	0.46	0.06	0.46
	(5)	0.52	0.06	0.52	0.50	0.04	0.50	0.47	0.05	0.47
	(6)	0.53	0.05	0.53	0.50	0.04	0.51	0.48	0.04	0.48

3

Spurious tail risk factors and asset prices

This work has been presented at the following conferences:

- ▷ *38th International Conference of the French Finance Association (AFFI)*, Rennes University & Rennes School of Business, Saint-Malo, France
- ▷ *6th International Workshop on Financial Markets and Nonlinear Dynamics (FMND)*, Paris, France

Abstract

In this paper, we argue that certain recent findings concerning the predictive ability of tail risk exposure, defined as the extremal dependence between asset returns and market returns, are likely spurious. We argue that these results are related to biases in the estimation procedure of the tail dependence coefficient (TDC) computed based on the joint behavior of equity returns, market returns, or other factors. Supported by a simulation framework, we show how this coefficient may capture a high level of correlation rather than tail dependence. Then, we replicate recent studies finding a relationship between crash risk exposure and future excess returns. We proceed to show that these results do not hold when we control for the correlation coefficient and other past return behavior.

Keywords: Asset pricing, Tail risk, Tail dependence, Copula, Extreme Value Theory
JEL Classification: C13, C53, G12, G17

3.1 Introduction

In the recent asset pricing literature, there has been increasing interest in discovering pricing anomalies related to compensation for bearing tail risks. Even more recently, some studies have extended the univariate tail risk concept to bivariate and multivariate cases, mainly to measure tail risk exposure to the market or a predefined set of factors. In this paper, after examining bias related to the estimation of these measures of exposure to stock market crashes, we show that the results identified in recent studies are most likely spurious.

Since the seminal work of Fama and French (1995, 1996), there has been extensive research attempting to identify additional factors explaining the cross-section of expected stock returns that the capital asset pricing model (CAPM) alone cannot explain (Sharpe, 1964). More specifically, there has been a growing stream of literature devoted to proving the existence of pricing anomalies related to compensation for bearing extramarket risks. This literature argues that investors are averse to downside losses and demand higher future excess returns for holding stocks with higher downside risk. In this sense, Ang et al. (2006) show that stocks that exhibit downside risk have higher future excess returns. Since then, many strategies have been used to identify the downside risk premium in the cross-section of expected returns.

Recent studies rely on extreme value theory (EVT) to measure tail risk as a proxy for the risk premium (Huang et al., 2012). The relationship between tail risk and future excess returns has been documented in recent papers, such as those of Lu and Murray (2019) and Atilgan et al. (2020), who find a negative return impact of a stock's univariate crash risk. In most cases, such studies concentrate on risk measures based on the univariate behavior of equity returns; thus, these studies mainly focus on stock crash risk alone in terms of crash probability. However, recently, a growing body of literature has focused on multivariate crash risk measures, which are mainly captured with the tail dependence coefficient (TDC). The TDC depicts the probability that extreme events involving several random variables occur simultaneously. In this paper, we focus on the TDC estimated based on the joint behavior of stock returns and market returns (Bali et al., 2014; Kelly and Jiang, 2014; Van Oordt and Zhou, 2016; Chabi-Yo et al., 2018) or the joint behavior of other factors (Chabi-Yo et al., 2021; Ruenzi et al., 2020). The idea to use the TDC as a proxy for crash aversion can be traced as far back as Poon et al. (2004), who consider the possibility of a premium for stocks that exhibit a TDC that increases with the market. The intuition behind such measures is that stocks sensitive to market crashes should include a premium.

Table 3.1 surveys six papers based on tail risk exposure variables related to the TDC between stock returns and market returns. These papers provide significant explanatory power for one-period-ahead stock returns. We report the name of the variable used in each study, its notation, the estimation procedure, the dataset, and the sample period of the study. Finally, the last column shows the main portfolio sorting results based on the tail risk measure in each study; it presents the high-minus-low (H – L) quintile portfolio and its associated t statistics. Kelly and Jiang (2014) construct a joint tail risk measure based on the TDC and the hill estimator developed by Hill (1975) and show that firms with higher tail risk exposure earn higher excess returns. Van Oordt and Zhou (2016) develop a crash risk measure called the tail beta (β^T), which helps predict losses

in future stock market crashes but does not incorporate a positive premium. Agarwal et al. (2017) develop a tail risk measure to explain hedge fund performance. He shows that hedge funds that exhibit higher tail risk exposure earn more excess returns. Meine et al. (2016) show that crash risk exposure is compensated by higher future excess returns in the cross-section of bank credit default swaps. Chabi-Yo et al. (2018) evaluate the crash risk exposure of an equity to the market with a copula approach and show that stocks with higher crash risk exposure earn higher future returns. Their results are confirmed with the same approach but in the international stock market by Weigert (2016). Chabi-Yo et al. (2021) extend the concept of crash risk exposure to the market to encompass crash risk exposure to risk factors, and their measure has a significantly positive effect on average future stock returns.

In this paper, we raise the following concerns regarding a well-known potential bias in the estimation of the TDC. Supported by a simulation framework, we reveal bias in both parametric and nonparametric TDC estimation procedures when data exhibit strong dependence as measured by the correlation coefficient. In the EVT literature, estimations of the TDC are normally conducted as a part of exploratory analyses in multivariate extremes, but this is usually a preliminary method used to determine whether data exhibit tail dependence and, hence, the type of models that might be suitable. Even these assessments, however, should be conducted only in conjunction with other dependence summaries because there are several models that exhibit no tail dependence (i.e., asymptotic independence), for which one would estimate quite large values of the TDC at any observable threshold (Frahm et al., 2005a). More generally, when the true underlying TDC is close to zero but there is dependence in the data, most estimators perform poorly (Poulin et al., 2007; Frahm et al., 2005a). Several studies have raised concerns regarding this bias and suggest that tail dependence should be tested before the TDC is estimated (Ledford and Tawn, 1996; Capéraà et al., 1997; Hoga, 2018). We show theoretically and in a simulation framework how such an estimation procedure could lead to biased estimations of the TDC.

Our research is also related to recent studies devoted to exposing biases in analyses of the cross-section of expected stock returns. Notably, Lewellen et al. (2010) show that the explanatory power of certain documented factors is most likely spurious. Harvey et al. (2016) argue that most of the relevant research findings in financial economics are likely false; they argue that such findings underemphasize the role of chance in the significance of the results. Harvey et al. (2016) also mention that in the fields of finance and economics, it is difficult to publish studies that replicate empirical findings related to traditional factors. Accordingly, he suggests that there is a bias toward publishing results related to new factors rather than rigorously testing the predictive ability of existing factors in multiple settings. This study is also related to a recent paper by McLean and Pontiff (2016), who argue that certain stock market anomalies become less anomalous after works identifying them are published.

The main results of this paper can be summarized as follows. First, we show how a crash risk exposure variable can be subject to potential bias. Second, we replicate common findings in research investigating crash risk exposure and find that this factor has a significant ability to predict future excess returns with impressive t statistics. Third, we observe that the effect and significance of these findings decrease and vanish when we control for the correlation with market returns. Fi-

nally, this study explains that bias is the main driver of the significance of the relationship between the examined crash exposure measure and future excess returns.

The rest of the paper is organized as follows. Section 3.2 presents the TDC and introduces the major concerns related to potential bias with respect to the estimation. Section 3.3 introduces the data and the variables used in the study. In Section 3.4, we present the results regarding the relationship between crash risk exposure as captured by several variables and future excess returns. We identify where bias could arise in the estimation and then control for it. We study the relationship between a spurious risk factor and future returns. Finally, Section 3.5 concludes.

3.2 The TDC and bias

In this section, we introduce the TDC and its link to the study of crash risk exposure. Then, we show the potential for bias in a naive estimation of this coefficient.

3.2.1 The TDC and the copula

The TDC depicts the probability that extreme events occur simultaneously in relation to several random variables. The TDC usually refers to the asymptotic probability introduced by Sibuya (1960) and subsequently defined by Joe (1997). This notion describes the dependence between extreme values in either the upper-right-quadrant tail or lower-left-quadrant tail of a bivariate distribution. A lower TDC, denoted λ, is defined as follows (Joe, 1997):

$$\lambda = \lim_{u \to 0^+} \mathbb{P}\left[X < F_X^{-1}(u) | Y < F_Y^{-1}(u)\right], \tag{1}$$

where $\lambda \in [0,1]$. We denote the generalized inverses of the univariate cumulative distribution functions F_X and F_Y as F_X^{-1} and F_Y^{-1}. The TDC is a pure copula property and is not based on marginal distributions but on only the copula, i.e., the marginal-free version of the joint distribution (Nelsen, 2007; Joe, 2014). The dependence structure is fully described by the copula function and holds independent of the marginal distributions. Let us denote as C the copula function between X and Y that can be expressed as follows:

$$C(u,v) = F\left(F_X^{-1}(u), F_Y^{-1}(v)\right), \tag{2}$$

where $(u,v) \in [0,1]^2$. The TDC is independent of the margins of X and Y; thus, we express the TDC solely in terms of the copula function as follows:

$$\lambda = \lim_{u \to 0^+} \frac{C(u,u)}{u}. \tag{3}$$

The generalization of the TDC in a multivariate setting with N factors is given as follows:

$$\lambda^X = \lim_{u \to 0^+} \mathbb{P}\left[X < F_X^{-1}(u) \mid \bigcup_{j=1}^{N} \left\{Y_j < F_{Y_j}^{-1}(u)\right\}\right]. \tag{4}$$

Table 3.1 : Major tail risk exposure studies.

Measure	Reference	Estimation	Dataset	Period	H-L spread
β^{T}	Van Oordt and Zhou (2016)	Nonparmetric (threshold $q = 0.04$) 1250 observations	Daily US returns NYSE, AMEX and NASDAQ	1963-2010	0.0% (-0.10)
CRASH	Chabi-Yo et al. (2018)	Parametric 250 observations	Daily US returns NYSE, AMEX and NASDAQ	1963-2012	4.32% (3.68)
CRASH	Ruenzi and Weigert (2018)	Parametric 250 observations	23 International equity markets	1963-2012	N.A.
CRASH	Weigert (2016)	Parametric 250 observations	$\approx$ 50,000 International stocks	1980-2014	11.54% (4.51) (US market)
Tail Risk	Agarwal et al. (2017)	Nonparametric (threshold $q = 0.05$) 500 observations	6,281 US equity funds	1996-2012	8.16% (2.16)
MCRASH	Chabi-Yo et al. (2021)	Nonparametric (threshold $q = 0.05$) 250 observations	Daily US returns NYSE, AMEX and NASDAQ	1964-2018	4.68% (3.69)

Notes: This table reports the main studies investigating the cross-section of tail risk exposure and future returns. We report the name of the variable, its notation, the corresponding reference, the estimation procedure, the dataset, and the study sample period. The last column shows the main results with respect to portfolio sorting based on the tail risk exposure measure in the corresponding study; we select the high-minus-low quintile portfolio and its associated t statistics.

The copula expression is given in appendix A.

3.2.2 Estimation of the TDC

The estimation of the TDC is often related to the estimation of copulas. Indeed, if one estimates a parametric copula, he or she can easily deduce the corresponding parametric TDC. Nonetheless, an accurate estimation of the TDC requires focusing merely on extreme observations. Such a parametric procedure, in which the whole dataset is used to estimate the copula function, may not be appropriate since it does not focus on the tail. To overcome the issue of choosing a specific parameterization of the copula function, some researchers have proposed a nonparametric version of the TDC estimator based on the empirical copula introduced by Deheuvels (1979). This estimator corresponds to a discretization of the TDC as defined by Joe (1997) and relies on the selection of a threshold over which the probability of the occurrence of joint extreme events is computed.

We consider n bivariate observations (x_j, y_j) for $j \in \{1, n\}$ generated with a dependence model of copula C. The nonparametric estimator of the lower TDC is defined as follows (Caillault and Guégan, 2005; Frahm et al., 2005a):

$$\widehat{\lambda}_L\left(\frac{i}{n}\right) = \frac{\widehat{C}\left(\frac{i}{n}, \frac{i}{n}\right)}{\frac{i}{n}}, \tag{5}$$

where $1 \geq i \geq n$, and $(u, v) \in [0,1]^2 \mapsto \widehat{C}(u, v)$ is the empirical copula introduced by Deheuvels (1979). We can write this empirical copula as follows (Genest and Rémillard, 2004):

$$\widehat{C}(u, v) = \frac{1}{n} \sum_{j=1}^{n} 1_{\{\widehat{F}_X(X_j) \leq u\}} 1_{\{\widehat{F}_Y(Y_j) \leq v\}}, \tag{6}$$

where $\widehat{F}_X$ and $\widehat{F}_Y$ are estimations of the marginal cumulative distribution functions. Focusing on X, $\widehat{F}_X$ is defined as follows:

$$\widehat{F_X}(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{X_j \leq x\}}. \tag{7}$$

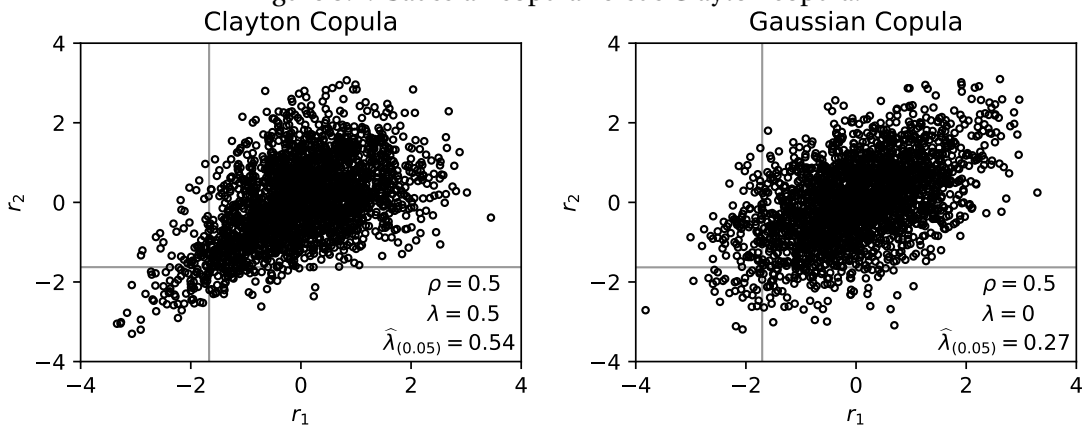
Schmidt and Stadtmüller (2006) show that a nonparametric estimator of the TDC exhibits strong consistency and is asymptotically normal.

The estimator of the lower TDC provided in Eq. (5) relies on the selection of an appropriate integer $i \in \{1, n\}$. Various selection rules for this free parameter have been proposed in the literature. An example is the plateau-finding algorithm (Frahm et al., 2005a). The selection rule for the threshold strongly impacts the quality of these nonparametric TDC estimators. Ideally, the threshold should allow us to focus on only a few extreme observations to prevent bias in the TDC estimation stemming from data from the bulk of the distribution. Nonetheless, the variance in the estimator would override.

3.2.3 Bias in the estimation of the TDC

Recent studies have documented bias linked to the estimation of the TDC. It is well established by extreme value theory (EVT) studies that the estimation of the TDC could result in misleading interpretations. Especially when the true underlying TDC is close to zero and the general dependence captured by the correlation coefficient is strong, most estimators perform poorly. This result is related to the findings by Frahm et al. (2005a) and Poulin et al. (2007), who suggest that tail dependence should be tested before an estimation of the TDC is carefully computed (Ledford and Tawn, 1996; Capéraà et al., 1997). A high level of correlation coefficients is often found in financial data and, therefore, might be captured by a TDC estimator.

Figure 3.1: Gaussian copula versus Clayton copula.



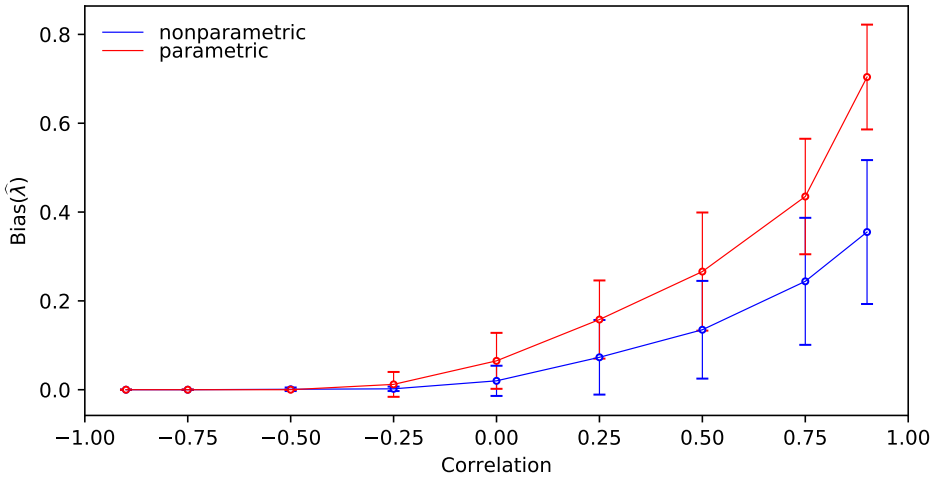
Notes: This figure displays 2500 random simulations of 2 bivariate distributions with Gaussian marginals. The graph on the left is drawn based on a Clayton copula with a theta parameter of 1 and an underlying TDC of 0.5. The graph on the right is drawn based on a Gaussian copula with a correlation coefficient of 0.5 and an underlying TDC of 0, as the Gaussian copula exhibits no tail dependence.

To graphically illustrate the potential for bias in estimations of the TDC, we show two examples in Fig. 3.1; in one example, the true underlying TDC is equal to 0.5, which is the Clayton copula, and in the other example, the true TDC is equal to 0, which is the Gaussian copula. A nonparametric estimation of the TDC in this example could lead to a bias of 0.27 in the Gaussian case. Frahm et al. (2005a) document similar results using a finite mixture of bivariate Gaussian distributions, although tail independent, which produce sample observations suggesting tail dependence even for large sample sizes.

A parametric method is also employed to estimate the TDC in Chabi-Yo et al. (2018); Ruenzi and Weigert (2018); Weigert (2016), as presented in appendix B. Chabi-Yo et al. (2018) attribute their findings to their parametric estimation procedure and explain that an estimation based on a whole joint distribution is more robust than one that relies on a small number of observations in the tail. To test this affirmation, we use a simulation framework to compare the nonparametric method with their parametric method. We compare the two estimators using random data samples generated by tail-independent bivariate distributions using a Gaussian copula function for $N = 50$ random sample replications with a sample size of $n = 250$. We use different correlation coefficient values $\rho \in \{-0.9, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 0.9\}$ to link the bias to the correlation coefficient. Then, we compute the average of the empirical bias and the standard deviation of the

sample bias. The results are presented in Fig. 3.2. The graph shows an upward bias with positive

Figure 3.2: Bias estimation in Gaussian simulations.



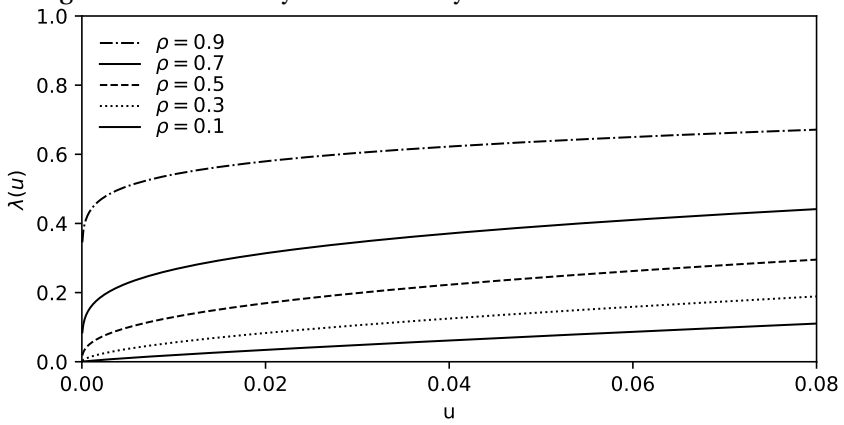
Notes: This figure shows the results of the bivariate Gaussian simulation framework with 20 samples generated with a sample size of $n = 250$ for different correlation coefficient values. The results are reported for two methods; the parametric method is shown in red, and the nonparametric method is shown in blue. The plain line represents the average level of bias computed for the 50 samples, and the error bar is the standard deviation of the bias.

values for the correlation coefficient. We can observe a nonlinear relationship between the bias in the estimator and the correlation coefficient. We also observe that the parametric method is more biased than the nonparametric method.

If we examine the limiting case in the Gaussian copula, although it exhibits a theoretical TDC at 0, we observe that the theoretical TDC increases as the correlation coefficient increases if we choose a threshold u in Eq. (3) slightly above 0. We plot these results in Fig. 3.3, which shows the expected bias of the estimation of the TDC with different threshold choices above the limiting case of $u = 0$ and different correlation coefficient values given that we face a bivariate Gaussian distribution. The relationship between the correlation coefficient and the TDC is largely hidden due to its strong nonlinear nature. In the simulation framework, the results suggest that bias increases nonlinearly as the correlation coefficient increases.

The corresponding bias behaves in the same manner for the multivariate TDC discussed in Chabi-Yo et al. (2021). As in the bivariate case, the nonparametric estimation of the multivariate TDC requires the selection of a specific threshold q. In Fig. 3.4, we plot the expected bias of the multivariate setting of equity returns and a 7-factor model (as elaborated in Chabi-Yo et al. (2021)). In this case, we vary only the correlation coefficient of the equity returns and the market factor MKT. We consider two cases. In the first case, the rest of the dependence between the returns and other factors is described by a correlation matrix with a null coefficient for every possible combination. We denote this correlation matrix $\mathbf{I_7}$. Second, we consider a case in which we estimate the rest of the correlation matrix of the factors as the average correlation matrix over the whole sample period. We denote this matrix $\widehat{\rho}_{\mathbf{X}}$, which varies only for the correlation coefficient between the returns and the first factor across its support. The procedure used to estimate the expected bias in the theoretical framework is given by replacing the multivariate empirical cop-

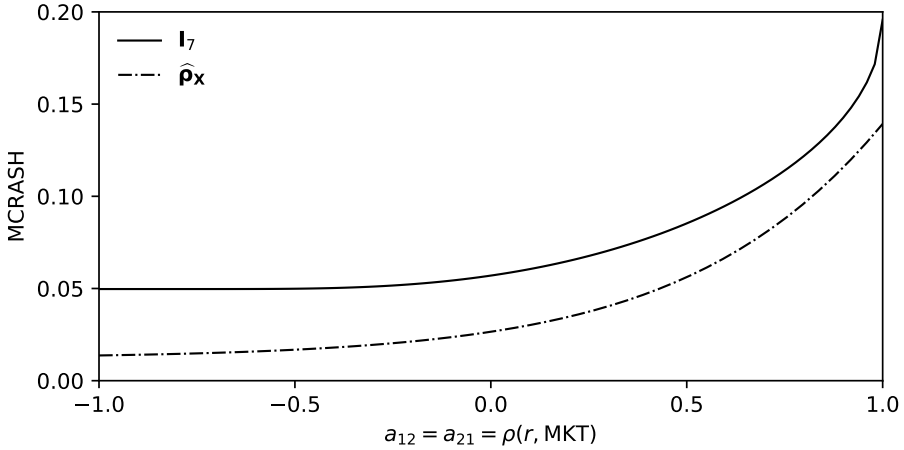
Figure 3.3: Sensitivity of the TDC by threshold selection and correlation.



Notes: This figure represents the value of the TDC in the Gaussian copula if it is not computed with the limiting case ($u \to 0$) but computed with u slightly above 0. The results are reported with different correlation coefficient values.

ula with a multivariate Gaussian copula with a given correlation matrix. The procedure used to obtain the corresponding copula expression is given in appendix A.

Figure 3.4: Multivariate TDC bias as a function of the correlation coefficient.



Notes: This figure shows the bias of the multivariate TDC in the multivariate Gaussian case with different correlation coefficients between the first two variables (in the case of multivariate crash risk [MCRASH], it corresponds to the correlation coefficient between returns and the first factor). The rest of the correlation matrix is computed with a null correlation coefficient for $\mathbf{I_7}$ and an estimation of the average correlation matrix between returns and other factors across the sample $\widehat{\rho}_{\mathbf{X}}$.

3.3 Data & variables

In Section 3.2, we argued that the TDC may be biased due to high correlation coefficients. Accordingly, the TDC might not capture how individual stocks behave during financial distress but rather the intensity of the dependence between stock returns and market returns. In this section, we present the stock dataset utilized and describe tail risk exposure and the other variables employed in our analysis.

3.3.1 Data

Our sample consists of data obtained from the Center for Research in Security Prices (CRSP) database (share code 10 or 11) corresponding to all common stocks traded on the NYSE, AMEX, and NASDAQ between January 1964 and December 2020. We select stocks with at least 200 nonzero return observations over the previous 250 trading days and prices of at least USD $2 to remain in the sample in month t. This procedure removes a large number of small and illiquid stocks from our sample. Thus, the sample is reduced to 1977989 observations corresponding to between 760 and 5937 firms in each month over the sample period. The risk-free rate and the factors MKT, SMB, HML, RMW, CMA, and UMD are derived from Kenneth French's website. The last factor (BAB) is downloaded from the AQR website[1].

3.3.2 Variables

The variables presented in Table 3.1 are the key variables used in this analysis; all the variables depend on the TDC introduced in Eq. (1) and may be subject to estimation bias and a spurious relationship with future excess returns. The crash measure (Chabi-Yo et al., 2018; Ruenzi and Weigert, 2018; Weigert, 2016) is defined as the TDC between the stock return and market returns as follows:

$$\text{CRASH}_i = \lim_{u \to 0+} P\left(r_i \leq F_i^{-1}(u) \mid r_m \leq F_m^{-1}(u)\right) = \hat{\lambda}_{im}(\text{u}), \tag{8}$$

where Fi and Fm denote the cumulative distribution functions of the stock return r_i and the market return r_m, respectively. The multivariate crash risk MCRASH (Chabi-Yo et al., 2021) is based on a generalization of the CRASH measure to the multivariate case. Chabi-Yo et al., 2021 define MCRASH as follows:

$$\text{MCRASH}_i = \lim_{u \to 0+} P\left(r_i \leq F_i^{-1}(u) \mid \bigcup_{j=1}^{N} \left\{X_j \leq F_j^{-1}(u)\right\}\right) = \hat{\lambda}_i^X(\text{u}), \tag{9}$$

where $\boldsymbol{X} = (X_1, \ldots, X_N)'$ denotes multiple factors. MCRASH is estimated using a seven-factor model in which the first five factors are those proposed by Fama and French (1995, 1996), i.e., the excess market return (MKT), the size factor (SMB), the value factor (HML), the profitability factor (RMW), and the investment factor (CMA). The other factors are the momentum factor (UMD) from Carhart (1997) and the betting-against-beta (BAB) factor from Frazzini and Pedersen (2014). The variable is estimated with its non-parametric estimator (the formula is given in the In the case of MCRASH, bias can arise from multiple correlation sources, but we focus on the correlation with the first factor MKT, which is also measured with $\rho(r_i, r_m)$.

The next variables depend on the TDC and are scaled based on the ratio of the univariate tail risk of stock i over the tail risk of the market. The tail beta β^{T} (Van Oordt and Zhou, 2016) is estimated nonparametrically with the following measure:

$$\hat{\beta}_i^T = \hat{\lambda}_{im}(\text{u})^{1/\tilde{\alpha}_m} \frac{\text{VaR}_i(u)}{\text{VaR}_m(u)}, \tag{10}$$

[1] `https://www.aqr.com/Insights/Datasets`

where the tail index $\hat{\alpha}_m$ is estimated with the Hill estimator (see Hill (1975)). The parameter (with range $(0,2)$) is used to determine the tail behavior such that the smaller the parameter is, the heavier the tail. In this study, it is estimated with a threshold of $q = 0.04$. Finally, the value at risk measures (VaRs) are the historical VaRs of the stock and the market estimated with the corresponding threshold u.

As mentioned in Section 3.2, the TDC may capture a high level of correlation dependence instead of a true asymptotic dependence. To confirm this relationship, we introduce the correlation coefficient between asset returns and market returns ($\rho(r_i, r_m)$). Usually, this relationship is captured through asset exposure to the market, which is the market beta of asset i (β_i) in the CAPM model. The market beta can be expressed as the scaled correlation to the level of idiosyncratic risk as follows:

$$\beta_i = \rho(r_i, r_m)\frac{\sigma(r_i)}{\sigma(r_m)}$$

where $\sigma(r_i)$ and $\sigma(r_m)$ are the respective volatilities. However, in this case, different levels of idiosyncratic risk ($\sigma(r_i)$) can lead to very different market beta values and may express different relationships. The relationships between the aforementioned variables and future returns may also be attributed to relationships with other attributes known to impact future equity returns. This is the case for cokurtosis (cokurt), with Fang and Lai (1997) and Dittmar (2002) showing that stocks with high cokurtosis earn high average returns. This is also the case for coskewness (coskew), which has already been shown to explain expected stock returns Harvey and Siddique (2000). The coskewness is given in Chabi-Yo et al. (2018) as follows:

$$\text{coskew} = \frac{E\left[\left(r_i - \mu_i\right)\left(r_m - \mu_m\right)^2\right]}{\sqrt{\text{VAR}\left(r_i\right)}\,\text{VAR}\left(r_m\right)},$$

Cokurtosis is given as follows:

$$\text{cokurt} = \frac{E\left[\left(r_i - \mu_i\right)\left(r_m - \mu_m\right)^3\right]}{\sqrt{\text{VAR}\left(r_i\right)}\,\text{VAR}\left(r_m\right)^{3/2}},$$

which are estimated nonparametrically. To control for univariate risk, we include the empirical value at risk of stock i (VaR ri) estimated at the 5% level over the previous 12 months, which is a common measure of left-tail risk. Following most studies that reduce the presence of outliers, we winsorize all the independent variables at the 1% level.

3.4 Empirical analysis

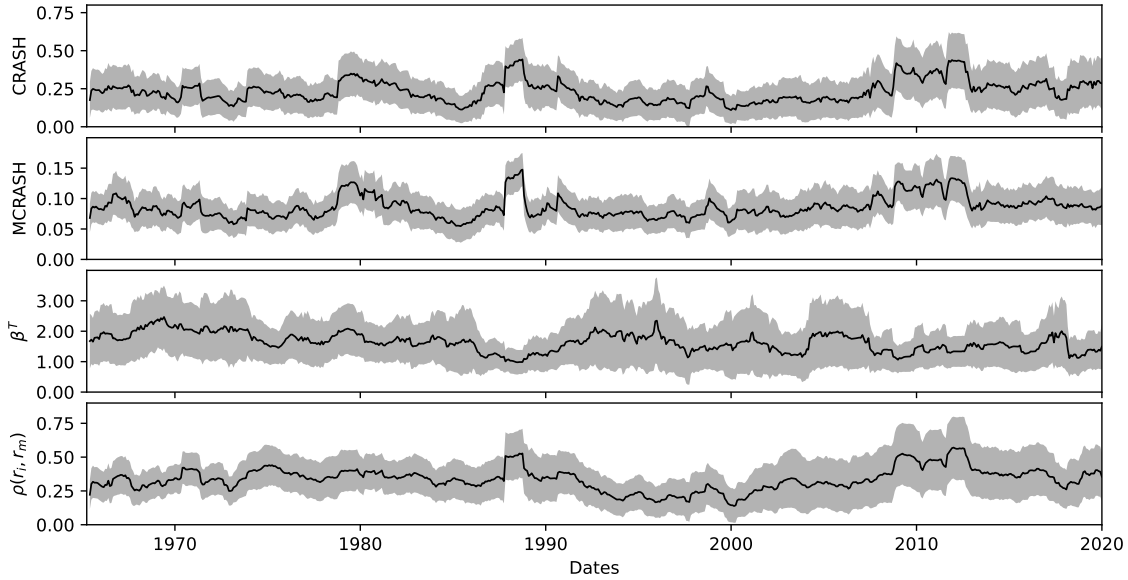
In Section 3.3, we describe the data and the variables used in this study. This section is dedicated to proving that the relationship between tail risk exposure and future excess returns is spurious. We first provide some descriptive statistics and attempt to reproduce the results previously obtained in the literature. Then, we show that the return predictability previously attributed to tail risk exposure is in fact driven by the correlation between stock and market returns and other past return behavior.

3.4.1 Descriptive statistics and temporal variation

The summary statistics of all the independent variables used in this analysis and their correlation matrix are provided in Table 3.2 . The table summarizes the results regarding tail risk exposure and the other variables introduced in Section 3.3. In panel A, we present the mean, standard deviation, skewness, kurtosis, minimum, 5th percentile, 25th percentile, median, 75th percentile, 95th percentile and maximum of each variable.

Panel B presents the correlation matrix of the independent variables. We observe positive correlation coefficients between the 3 risk exposure variables. The correlation coefficient between CRASH and MCRASH is relatively high at 0.72, indicating a strong relationship. The lowest correlation coefficients are those of β^T and MCRASH at 0.18 and β^T and CRASH at 0.26. Regarding the correlation coefficient between $\rho(r_i, r_m)$ and the 3 variables, we observe positive values ranging from 0.07 to 0.73. The correlations of CRASH and MCRASH with $\rho(r_i, r_m)$ are relatively high, with values of 0.73 and 0.54, respectively; however, we observe that the dependence is far from perfect, suggesting that the variables may measure different types of exposure to market returns. We also find positive relationships between these 3 variables and the market beta, with values that range from 0.20 to 0.46.

Figure 3.5: Aggregated behavior of tail risk exposure and return correlation over time.



Notes: This figure plots the aggregate behavior of the tail risk measure introduced in section 3.1 over the sample period. The bold line refers to the monthly cross-sectional average of each variable, and the gray area represents the standard deviation interval around the average.

Fig. 3.5 presents an overview of the aggregated measures across the sample period. No clear temporal trend appears in the aggregated behavior of any variable. However, we can observe jumps in CRASH and MCRASH that coincide with jumps in the average correlation with the market. These spikes in the aggregated behavior seem to be related to major financial crisis events (mainly Black Monday in 1987, the subprime crisis and the relatively recent COVID-19 crash at the beginning of 2020). Increases in this correlation during periods of financial distress have been

Table 3.2 : Summary statistics and correlations.

Panel A: Summary statistics

	Mean	StDev.	skew	kurt	min	q5	q25	med	q75	q95	max
crash	0.23	0.15	0.63	-0.14	0.00	0.08	0.00	0.24	0.32	0.55	0.64
mcrash	0.09	0.03	0.33	-0.26	0.02	0.06	0.03	0.08	0.11	0.15	0.18
β^{T}	1.59	0.91	0.51	0.45	0.00	1.04	0.00	1.48	2.09	3.28	4.31
β^{MKT}	0.99	0.55	0.24	0.24	-0.28	0.65	0.08	0.98	1.31	1.95	2.54
β^{SMB}	0.69	0.72	0.46	0.17	-0.89	0.17	-0.38	0.63	1.13	1.99	2.80
β^{HML}	0.16	0.82	-0.36	1.04	-2.44	-0.27	-1.28	0.19	0.64	1.46	2.30
β^{UMD}	-0.04	0.59	-0.07	1.27	-1.85	-0.35	-1.04	-0.03	0.28	0.93	1.73
$\rho(R_i, R_m)$	0.33	0.19	0.22	-0.65	-0.05	0.18	0.03	0.32	0.47	0.67	0.78
coskew	-0.09	0.24	-1.57	6.13	-1.25	-0.18	-0.46	-0.07	0.04	0.24	0.43
cokurt	1.75	1.66	3.16	15.62	-0.36	0.78	0.07	1.41	2.26	4.24	12.18

Panel B: correlation

	CRASH	MCRASH	β^{T}	β^{MKT}	β^{SMB}	β^{HML}	β^{UMD}	$\rho(R_i, R_m)$	coskew	cokurt
CRASH	1.00									
MCRASH	0.72	1.00								
β^{T}	0.25	0.18	1.00							
β^{MKT}	0.30	0.20	0.46	1.00						
β^{SMB}	-0.05	0.06	0.32	0.40	1.00					
β^{HML}	-0.00	0.05	-0.10	0.12	0.16	1.00				
β^{UMD}	0.04	0.10	0.00	-0.08	-0.05	0.03	1.00			
$\rho(R_i, R_m)$	0.73	0.54	0.07	0.44	-0.07	-0.03	0.00	1.00		
coskew	-0.29	-0.26	0.01	-0.04	-0.05	-0.04	-0.05	-0.06	1.00	
cokurt	0.57	0.45	-0.02	0.27	-0.05	-0.02	0.03	0.63	-0.53	1.00

Notes: Panel A reports the summary statistics of the different tail risk exposure measures introduced in Section 3.3 and the other variables used in the analysis. The results regarding the three different tail risk measures (CRASH, MCRASH, and β^{T}) are given; we also include the betas from the Carhart 4-factor model (Carhart, 1997) (β^{MKT}, β^{SMB}, β^{HML}, β^{UMD}), the correlation coefficient between asset returns and market returns ($\rho(r_i, r_m)$), coskewness (coskew), and cokurtosis (cokurt). In the columns, we report the mean, standard deviation (StDev), skewness (skew), kurtosis (kurt), minimum value (min), 5% quantile (q5), 25% quantile (q25), 50% quantile (median), 75% quantile (q75), 95% quantile and maximum value (max) of each variable. We select all the US stocks with CRSP share codes 10 and 11 traded on the NYSE/AMEX/NASDAQ between January 1965 and January 2020, excluding stocks with prices below $2 on the portfolio formation date. We also require that a stock have at least 200 nonzero return observations over the previous year to remain in the sample. In panel B of this table, we report the correlation matrix of the variables introduced in panel A.

well documented.

3.4.2 Portfolio sorts

Following Daniel and Titman (1997), we apply a double-sorting procedure. We conduct bivariate portfolio sorting with five portfolios based on the tail risk measure; then, within each quintile portfolio, we form five new portfolios based on the correlation coefficient between asset returns and market returns. We evaluate average returns in excess of the risk-free rate over month $t+1$ (i.e., returns minus the risk-free rate) for each of the 25 double-sorted portfolios. The results are reported in Table 3.3 . In column (H – L), we display the differences in returns between the lowest quantile portfolio (Low) and the highest quantile portfolio (High). We report the significance of this spread in the last column with the t statistics of the test computed with Newey et al. (1987) standard errors with 12 lags. We conduct this sorting for each of the 3 tail risk exposure indicators, with the results for CRASH, MCRASH, and β^T presented in panels A to C, respectively.

First, considering only univariate sorting based on tail risk exposure (provided in the first line of each panel with the index All), we observe a significant excess returns spread between the portfolios in the lowest quintile and those in the highest quintile as the tail risk measure increases. The spread is significant at the 1% level, with high t statistics for each of the tail risk exposure measures in the 3 panels. We observe a positive spread of the CRASH and MCRASH measures. In panel A, the CRASH return spread (H – L) amounts to 0.58% per month with a t statistic of 16.48, and the stocks in the highest quantile earn 6.95% higher annualized excess returns. In the MCRASH return spread, we observe 0.54% per month with a t statistic of 14.64, which corresponds to 6.48% in annualized returns. In contrast, we observe a negative spread of -0.72% for the tail beta with a t statistic of -16.44.

However, considering this spread within the sorted portfolios based on the correlation values (rows Low to High), we observe that this spread is in fact sensitive to differences in the correlation values. The predictive power of the differences in the stock crash risk measures appears to be linked to different values of the correlation coefficient. Specifically, except for the nonsignificant spread in panel A, an increase in the correlation coefficient tends to accompany an increase in the return spread. More specifically, in panel B, in the lowest correlation quantile portfolio, the spread is significantly negative, while it is significantly positive in the highest quantile correlation portfolio. The lowest correlation quintile portfolio exhibits an H – L MCRASH return spread of -0.58% (with t statistic -3.36), while the highest correlation quintile portfolio has a spread of 0.74% (t statistic 6.29). In panel C, the lowest correlation quintile portfolio has a return spread of -0.65% (t statistic -6.16), while the highest a spread of -0.21% (t statistic -2.11). Consequently, in this setting, higher average future returns associated with higher tail risk exposure are driven by different values of the correlation coefficient.

3.4.3 Multivariate analysis

To test the relationships between the tail risk exposure measures and future excess returns, we proceed with Fama and MacBeth (1973) regressions at the individual firm level. Thus, we imple-

Table 3.3 : Equally weighted portfolio sorts based on correlation and tail risk.

	Low	2	3	4	High	H – L	t stat
Panel A: CRASH and correlation							
All	0.505	0.473	0.479	0.585	1.088	0.583^{***}	16.483
Low ρ_{r_i,r_m}	0.473	0.500	0.526	0.437	0.476	0.003	0.006
2	0.389	0.263	0.222	0.175	0.054	-0.335^{*}	-1.844
3	0.558	0.463	0.482	0.489	0.587	0.029	0.282
4	0.926	0.767	0.686	0.726	0.823	-0.104	-1.000
High ρ_{r_i,r_m}	1.674	1.036	0.623	0.842	1.311	-0.363^{*}	-1.797
Panel B: MCRASH and correlation							
All	0.559	0.289	0.542	0.591	1.106	0.547^{***}	14.638
Low ρ_{r_i,r_m}	0.745	0.229	0.345	0.219	0.160	-0.585^{***}	-3.356
2	0.394	0.065	0.385	0.251	0.047	-0.347^{***}	-2.692
3	0.362	0.389	0.568	0.421	0.665	0.303^{***}	3.237
4	0.594	0.535	0.707	0.739	1.071	0.478^{***}	5.565
High ρ_{r_i,r_m}	0.708	0.451	0.779	0.977	1.452	0.744^{***}	6.286
Panel C: β^T and correlation							
All	0.662	0.803	0.846	0.780	-0.054	-0.716^{***}	-16.436
Low ρ_{r_i,r_m}	0.593	0.641	0.654	0.458	-0.056	-0.649^{***}	-6.160
2	0.521	0.527	0.586	0.377	-0.555	-1.076^{***}	-11.567
3	0.659	0.698	0.713	0.624	-0.284	-0.943^{***}	-10.818
4	0.861	0.799	0.811	0.955	0.353	-0.508^{***}	-6.132
High ρ_{r_i,r_m}	0.965	1.100	1.231	1.380	0.759	-0.206^{**}	-2.114

Notes: This table reports the average monthly excess return of equally weighted bivariate portfolio sorts on the basis of the correlation coefficient and the corresponding tail risk measure. First, the stocks are sorted into quintile portfolios based on the tail risk measure estimated over the previous 12 months; then, within each quintile portfolio, we form five new portfolios based on the correlation coefficient between asset returns and market returns estimated over the previous 12 months. We select NYSE/AMEX/NASDAQ stocks with CRSP share codes 10 and 11 traded between January 1965 and January 2020, excluding stocks with prices below $2 on the portfolio formation date. We also require that a stock have at least 200 nonzero return observations over the previous year to remain in the sample. The results are divided into 3 sets, with one set for each of the following different tail risk measures: CRASH, MCRASH, and β^T. The H – L columns report the results of the high-minus-low portfolios, and the last column reports the associated t statistics with Newey et al. (1987) standard errors based on twelve lags. The superscripts ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

Table 3.4 : Fama and MacBeth (1973) regressions.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: CRASH							
CRASH	0.752**	1.119***	1.114***	1.298***	1.253***	1.133***	0.650***
β^{MKT}		-0.255*	-0.274**	-0.308**	-0.271**	-0.264*	-0.419**
β^{SMB}			0.055	0.054	0.039	0.038	0.119
β^{HML}				0.112	0.189**	0.181**	0.191**
β^{UMD}					-0.008	0.007	-0.004
coskew						1.133***	-0.146
cokurt							0.262**
Const.	0.395*	0.583***	0.568***	0.555***	0.536***	0.497***	0.395*
Panel B: MCRASH							
MCRASH	3.630***	4.611***	4.549***	5.021***	4.648***	3.983***	2.660***
β^{MKT}		-0.221*	-0.204**	-0.226*	-0.191	-0.203	-0.399**
β^{SMB}			0.002	-0.003	-0.022	-0.017	0.092
β^{HML}				0.103	0.183**	0.173**	0.185*
β^{UMD}					-0.018	-0.013	-0.018
coskew						3.983***	-0.207
cokurt							0.284**
Const.	0.254	0.400*	0.396**	0.367*	0.373*	0.350*	0.254
Panel C: β^T							
β^T	-0.176	-0.188	-0.205*	-0.202*	-0.266***	-0.299***	-0.246***
β^{MKT}		0.001	0.036	0.03	0.124	0.109	-0.145
β^{SMB}			0.007	-0.002	-0.004	0.009	0.104
β^{HML}				0.058	0.121	0.107	0.131
β^{UMD}					0.103	0.092	0.064
coskew						-0.299***	-0.576***
cokurt							0.285***
Const.	0.917***	0.908***	0.890***	0.866***	0.852***	0.770***	0.917***

Notes: This table reports the results of cross-sectional Fama and MacBeth (1973) regressions of future excess returns on the different tail risk measures (CRASH, MCRASH, and β^T). We control for the market beta β^{MKT}, small-minus-big beta β^{SMB}, high-minus-low beta β^{HML}, coskewness and cokurtosis. We select NYSE/AMEX/NASDAQ stocks with CRSP share codes 10 and 11 from between January 1965 and January 2020, excluding stocks with prices below \$5 on the portfolio formation date. We also require that a stock have at least 200 nonzero return observations over the previous year to remain in the sample. The results are divided into 3 panels, each corresponding to one of the following tail risk measures: CRASH, MCRASH, and β^T. The significance of the coefficient is given using the t statistics with Newey et al. (1987) standard errors based on twelve lags. The superscripts ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

ment the following model:

$$R^e_{it+1} = \alpha_0 + \alpha_1^{TailM}.TailM_{i,t} + \sum_{j=1}^{K} \alpha_j Y^j_{i,t} + \varepsilon_{it+1}, \tag{11}$$

where $TailM$ is the corresponding tail risk exposure variable considered, and $Y^j_{i,t}$ denotes the set of control variables used. Table 3.4 reports the results of the models corresponding to each of the 3 tail exposure measures in 3 different panels. In models (1) to (7), we successively add the betas from the Carhart 4-factor model (Carhart, 1997) β^{MKT}, β^{SMB}, β^{HML}, β^{UMD}, coskewness and cokurtosis. The significance of each coefficient is given using t statistics with Newey et al. (1987) standard errors based on twelve lags.

Consistent with the results of previous studies, we find positive values for the coefficients of CRASH and MCRASH with similar magnitudes (Chabi-Yo et al., 2018, 2021). Additionally, for the β^T coefficients, which are also shown to be related in a study (Van Oordt and Zhou, 2016), we find negative values. In panel A, we find that the impact of CRASH is statistically significant at the 1% level, except for in the first model (1), where it is significant at the 5% level, with coefficient values ranging from 0.65% to 1.30%. In panel B, the MCRASH measure is significant at the 1% level across all models (1) to (7) with coefficients ranging from 3.63% to 5.02%. In panel C, the coefficients range from -0.30% to -0.18% but are significant at the 1% level in only the last three models (models (5), (6) and (7)) and significant at the 10% level in models (2) and (3).

3.4.4 Additional controls

Table 3.5 : Fama and MacBeth (1973) regressions and quintile portfolio on correlation.

	All	Low	(2)	(3)	(4)	High
Panel A: CRASH						
CRASH	0.348*	0.234	0.159	0.278	0.341	0.248
	(1.95)	(0.29)	(0.38)	(0.84)	(1.36)	(0.57)
past ret	0.655**	1.234**	0.524	0.086	1.014**	0.010
	(2.44)	(2.27)	(1.41)	(0.21)	(2.24)	(0.01)
VaR_{r_i}	0.355***	0.438***	0.268***	0.319***	0.325***	0.638**
	(6.39)	(5.07)	(3.45)	(4.5)	(3.85)	(2.53)
Panel B: MCRASH						
MCRASH	1.066*	1.176	2.846**	-0.688	1.664*	1.629
	(1.77)	(0.66)	(2.28)	(-0.50)	(1.68)	(0.73)
past ret	0.651**	1.170**	0.519	0.101	0.992**	-0.611
	(2.43)	(2.17)	(1.43)	(0.25)	(2.21)	(-0.79)
VaR_{r_i}	0.352***	0.424***	0.265***	0.318***	0.314***	0.445***
	(6.35)	(4.68)	(3.40)	(4.47)	(3.72)	(2.62)
Panel C: β^T						
β^T	0.105*	0.080	-0.068	0.048	0.209	0.413*
	(1.91)	(0.50)	(-0.65)	(0.43)	(1.55)	(1.90)
past ret	0.654**	1.216**	0.495	0.112	1.022**	0.104
	(2.43)	(2.25)	(1.32)	(0.27)	(2.27)	(0.14)
VaR_{r_i}	0.383***	0.434***	0.248***	0.325***	0.393***	0.838***
	(6.65)	(4.97)	(3.11)	(3.85)	(3.68)	(2.94)

Notes: This table reports the results of cross-sectional Fama and MacBeth (1973) regressions of future excess returns on the different tail risk measures (CRASH, MCRASH, and β^T). The control variables correspond to equation (7) in Table 3.4 ; we add controls for past return behavior (past ret) and the value at risk at the 5% level (VaR). We give the results for the whole dataset in the first column (All); then, we sort the stocks in our sample into quintile portfolios according to the correlation coefficient. We select NYSE/AMEX/NASDAQ stocks with CRSP share codes 10 and 11 traded between January 1965 and January 2020, excluding stocks with prices below $5 on the portfolio formation date. We also require that a stock have at least 200 nonzero return observations over the previous year to remain in the sample. The results are divided into 3 panels, each corresponding to one of the following tail risk measures: CRASH, MCRASH, and β^T. The significance of the coefficient is given using t statistics with Newey et al. (1987) standard errors based on twelve lags. The superscripts ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

We address previous concerns that the significant predictive ability of the different tail risk exposure variables is related to a confounding effect with the correlation coefficient. We investigate

the validity of the previous Fama and MacBeth (1973) regression results in Table 3.4 . We start with model (7) from Table 3.4 , where we add controls for past return behavior (past ret) and the value at risk at the 5% level (VaR) to control for the univariate tail risk and determine whether the effect of tail risk exposure still holds. We then proceed with a regression analysis of the quintile portfolios formed at the correlation coefficient level.

The results are reported in Table 3.5 . We give the results of the whole dataset in the first column (All); then, we sort the stocks in our sample into quintile portfolios according to the correlation coefficient in the next 5 columns. In panel A, considering the whole dataset (column All), we find that the coefficient is significant at only the 10% level with a value of 0.35%. Past returns and the value at risk are significant at the 5% and 1% levels with values of 0.65% and 0.35%, respectively. When considering the quintile portfolios (from columns Low to High), we find that the coefficients are no longer significant, with values ranging from 0.16 to 0.34.

In panel B, the coefficient of the entire dataset (column All) is still significant at the 10% level with a value of 1.07%. However, when we consider the quintile portfolios, the coefficient is no longer significant in the Low, (3), or High columns. The coefficient is only significant in the (2) column at the 5% level and in the (4) column at the 10% level. We observe results for the past returns and value at risk coefficients that are similar to those in panel A.

In the last panel, the coefficient becomes positive when the variables past ret and value at risk are added as controls, in contrast to the previous regressions. Additionally, in all the different quintile portfolios, the coefficients are no longer significant.

3.5 Conclusion

A plethora of factors have been recently introduced in the asset pricing literature, it is famously referred to as a "zoo of new factors" by Cochrane (2011). This has given rise to a debate regarding how many of these factors are really useful in providing independent information. In fact, some studies argue that most of the claimed results may be attributed to data mining (Harvey et al., 2016) or spurious regression (Deng, 2014). In this paper, we contribute to this ongoing debate by re-examining the findings of recent distress risk factors that explain the cross-section of expected stock returns. Specifically, we focus on three variables that capture the crash sensitivity of a stock to the market or multiple factors, measured with the TDC. In the corresponding studies, the authors claim that stocks sensitive to market crashes should include a premium or may be useful to predict future excess returns.

We first highlight possible bias in the estimation of the TDC when there is a high level of correlation between the corresponding random variables. Consequently, this bias may pollute all crash sensitivity measures and downside risk measures in empirical studies. Then, we replicate the recent studies finding a relationship between crash risk exposure and future excess returns. We show that these results do not hold when we control for the correlation coefficient and other past return behavior. The pricing anomalies are most likely driven by a confounding effect with the correlation coefficient. Indeed, the TDC might not capture how individual stocks behave during financial distress but rather the intensity of the dependence with the market.

Our paper contributes to the literature dedicated to revealing biases in the context of asset pricing. As mentioned by Harvey et al. (2016), there is a bias toward publishing studies investigating new factors. We suggest that more rigorous testing and estimation should be applied to empirical findings related to traditional factors. More specifically, our results call for careful analysis in empirical studies involving the estimation of the TDC with financial returns. Other measures to capture the tail risk exposure such as the marginal expected shortfall (Idier et al., 2014) should be considered in future works.

3.1 Multivariate TDC in terms of copula functions

Here, we detail the expression of MCRASH in terms of copula functions. First, given the MCRASH definition

$$\text{MCRASH}_i^X = \lim_{u\to 0^+} \mathbb{P}\left[X < F_X^{-1}(u) \mid \bigcup_{j=1}^{N}\left\{Y_j < F_{Y_j}^{-1}(u)\right\}\right], \tag{A.1}$$

we can express the equation in the following form:

$$\text{MCRASH}_i^X = \frac{\mathbb{P}\left(U_i < q \cap \bigcup_{j=1}^{N}\{U_j < q\}\right)}{\mathbb{P}\left(\bigcup_{j=1}^{N}\{U_j < q\}\right)}, \tag{A.2}$$

where $U_i = F_i(X_i)$, which gives the following nonparametric estimator:

$$\widehat{\text{MCRASH}}_i^X = \frac{\sum_{s=1}^{N} \mathbf{1}_{\{U_{i,s}\le q\}} \cdot \mathbf{1}_{\{U_{j,s}\le q \text{ or ... or } U_{j,s}\le q\}}}{\sum_{s=1}^{N} \mathbf{1}_{\{U_{j,s}\le q \text{ or ... or } U_{j,s}\le q\}}}. \tag{A.3}$$

In the case of independence, the estimator has a fixed bias of $\mathbb{P}\left(U_i < q\right) = q$.

To express MCRASH in terms of copula functions, we first provide the following expression of the denominator of Eq. (A.2):

$$\begin{aligned}
\mathbb{P}\left(U_i < q \cap \bigcup_{j\neq i} U_j < q\right) &= \mathbb{P}\left[\left(U_1 < q \cap U_2 < q\right) \cup \left(U_1 < q \cap U_3 < q\right) \cup \cdots \cup \left(U_1 < q \cap U_n < q\right)\right] \\
&= \sum_{k=2}^{n} \mathbb{P}\left(U_1 < q, U_k < q\right) - \sum_{2\le k_1 \leqslant k_2 \leqslant n}^{n} P\left(U_1 < q, U_{k_1} < q, U_{k_2} < q\right) \\
&\quad + \sum_{2\le k_1 \le k_2 \le k_3 \le n}^{n} \mathbb{P}\left(U_1 < q, U_{k_1} < q, U_{k_2} < q, U_{k_3} < q\right) - \ldots \\
&\quad + (-1)^{n-1}\mathbb{P}\left(U_1 < q, U_2 < q, \ldots, U_n < q\right) \\
&= \sum_{k=2}^{n} C_{1,k}(q,q) - \sum_{2\le k_1 \le k_2 \le n}^{n} C_{1,k_1,k_2}(q,q,q) + \sum_{2\le k_1 \le k_2 \le k_3 \le n}^{n} C_{1,k_1,k_2,k_3}(q,q,q,q) \\
&\quad - \sum_{2\le k_1 \le \ldots \le k_4 \le n}^{n} C_{1,k_1,\ldots,k_4}(q,q,q,q,q) + \ldots + (-1)^{n-1} C_{1,2,\ldots,n}(q,\ldots,q).
\end{aligned} \tag{A.4}$$

Then, we give the numerator as follows:

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{j\neq i} U_j < q\right) &= \mathbb{P}\left(U_2 < q \cup U_3 < q \cup \cdots \cup U_n < q\right) \\
&= \sum_{k=2}^{n} \mathbb{P}\left(U_k < q\right) - \sum_{2\le k_1 \leqslant k_2 \leqslant n}^{n} \mathbb{P}\left(U_{k_1} < q, U_{k_2} < q\right) \\
&\quad + \sum_{2\le k_1 \le k_2 \le k_3 \le n}^{n} \mathbb{P}\left(U_{k_1} < q, U_{k_2} < q, U_{k_3} < q\right) - \ldots \\
&\quad + (-1)^{n-1}\mathbb{P}\left(U_2 < q, \ldots, U_n < q\right) \\
&= \sum_{k=2}^{n} q - \sum_{2\le k_1 \le k_2 \le n}^{n} C_{k_1,k_2}(q,q) + \sum_{2\le k_1 \le k_2 \le k_3 \le n}^{n} C_{k_1,k_2,k_3}(q,q,q) \\
&\quad - \sum_{2\le k_1 \le \ldots \le k_4 \le n}^{n} C_{k_1,\ldots,k_4}(q,q,q,q) + \ldots + (-1)^{n-1} C_{2,\ldots,n}(q,\ldots,q).
\end{aligned} \tag{A.5}$$

3.2 Appendix simulation studies and estimation

The estimation procedure consists of combining copulas with various tail behaviors that exhibit no tail dependence (the Gauss, Frank, Farlie–Gumbel–Morgenstern [FGM], and Plackett copulas), lower-tail dependence (the Clayton, rotated Gumbel, rotated Joe, and rotated Galambos copulas), and upper-tail dependence (the Gumbel, Joe, Galambos, and rotated Clayton copulas). We consider all $4 \times 4 \times 4 = 64$ possible combinations with one lower tail-dependent copula, C_{LTD}; one copula that is asymptotically independent, C_{NTD}; and one copula that allows for asymptotic dependence in the upper tail, C_{UTD}:

$$C(u_1, u_2, \Theta) = w_1 \times C_{\text{LTD}}(u_1, u_2; \theta_1) + w_2 \times C_{\text{NTD}}(u_1, u_2; \theta_2) + (1 - w_1 - w_2) \times C_{\text{UTD}}(u_1, u_2; \theta_3), \quad \text{(B.1)}$$

where Θ denotes the set of the basic copula parameters $\theta_i, i = 1, 2, 3$ and the weights w_1 and w_2.[11]. The method consists of estimating every set of parameters Θ_j for $j = 1, \ldots, 64$ different copulas $C_j\left(\cdot, \cdot; \Theta_j\right)$ and then choosing the appropriate combination that minimizes the distance to the empirical copula. The Python package pycop (Nicolas, 2022) was used for generating data from normal copula functions and for the estimation of the combinations of copula functions.

Conclusion

The first part of this dissertation focuses on understanding how risk emerges in financial markets. To do so, we use sentiment analysis and text mining techniques to derive a sentiment proxy from weekly aggregations of online messages posted on the social media platform StockTwits. We employ numerical methods to estimate the parameters of a model of opinion formation. Consistent with previous research that has found that volatility is driven by herding behavior, we investigate the relationship between herding intensity and the level of volatility. In particular, we find that herding behavior was significantly higher and played a major role in the sentiment formation process regarding cryptocurrencies during the bubble period.

In future research, we plan to extend our analysis of the relationship between economic herding intensity and future stock returns behavior. We will also explore whether the estimated contagion parameters can serve as early warning signals for potential stock market risks. This will provide valuable insights into the potential uses of sentiment analysis and text mining techniques for identifying and mitigating financial risks.

In the second essay, we introduce a new statistical method for estimating the extremal dependence between two random variables. This method is based on the well-known tail dependence coefficient (TDC), for which there is no theoretical basis for selecting a threshold. In its nonparametric version, the estimation of the TDC requires the choice of an arbitrary threshold above which the probability of observing joint extreme values must be calculated. The main contribution of this paper is the proposal of a theoretical framework for selecting this threshold. The performance of the estimator is then evaluated through simulations and compared to the estimators used in traditional approaches. The results show the consistency of the proposed new estimator, although it does not clearly outperform the other estimators. The estimation method is then applied to evaluate the TDC between the returns of the US equity market index and the returns of the equity market indexes of 17 developed countries.

Based on these findings, we plan to conduct further research to incorporate the TDC into a portfolio optimization program. This will allow us to combine financial assets that are less likely to experience simultaneous crashes. We will test this technique during a global stock market crash to evaluate its effectiveness in mitigating financial risks. By providing a more accurate and robust

measure of extremal dependence, the TDC-based portfolio optimization program has the potential to improve the efficiency and stability of financial portfolios. We believe that this research will contribute to the development of more effective risk management strategies in finance.

The third chapter focuses on the relationship between crash sensitivity and future excess returns. Crash sensitivity is defined as the tail dependence between the returns of a financial asset and the returns of the market. It measures the probability that one asset will experience an extreme event, given that the market has also experienced an extreme event. Using a simulation framework, we demonstrate the bias in both the parametric and nonparametric TDC estimation procedures when the data exhibit strong dependence (as measured by the correlation coefficient). Finally, we replicate recent studies that explain future excess returns using crash risk sensitivity. However, we find that these results do not hold when we control for the correlation coefficient and other past return behavior. This suggests that crash sensitivity alone may not be a sufficient predictor of future excess returns, and that other factors should be taken into account when modeling financial risk.

As part of an ongoing research, we are exploring the use of other tail dependence measures as proxies for crash sensitivity. One such measure is the marginal expected shortfall, which measures a firm's expected returns when the market falls below a certain threshold. Unlike other measures, the marginal expected shortfall is not prone to statistical biases. In future studies, we will evaluate the performance of the marginal expected shortfall and other tail dependence measures in predicting crash sensitivity and future excess returns. We will also compare these measures to existing methods, such as the parametric and nonparametric TDC estimators discussed in the previous chapter. By improving our understanding of the link between crash sensitivity and future returns, we hope to contribute to the development of more effective risk management strategies in finance.

Bibliography

Abdous, B. and Theodorescu, R. (1992). Note on the spatial quantile of a random vector. *Statistics and probability letters*, 13(4):333–336.

Agarwal, V., Ruenzi, S., and Weigert, F. (2017). Tail risk in hedge funds: A unique view from portfolio holdings. *Journal of Financial Economics*, 125(3):610–636.

Aghakouchak, A., Ciach, G., and Habib, E. (2010). Estimation of tail dependence coefficient in rainfall accumulation fields. *Advances in water resources*, 33(9):1142–1149.

Alfarano, S., Lux, T., and Wagner, F. (2005). Estimation of agent-based models: the case of an asymmetric herding model. *Computational Economics*, 26(1):19–49.

Alfarano, S., Lux, T., and Wagner, F. (2008). Time variation of higher moments in a financial market with heterogeneous agents: An analytical approach. *Journal of Economic Dynamics and Control*, 32(1):101–136.

Alfarano, S. and Milaković, M. (2009). Network structure and n-dependence in agent-based herding models. *Journal of Economic Dynamics and Control*, 33(1):78–92.

Ang, A., Chen, J., and Xing, Y. (2006). Downside risk. *The review of financial studies*, 19(4):1191–1239.

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of finance*, 59(3):1259–1294.

Atilgan, Y., Bali, T. G., Demirtas, K. O., and Gunaydin, A. D. (2020). Left-tail momentum: Underreaction to bad news, costly arbitrage and equity returns. *Journal of Financial Economics*, 135(3):725–753.

Baker, M. and Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of economic perspectives*, 21(2):129–152.

Bali, T. G., Cakici, N., and Whitelaw, R. F. (2014). Hybrid tail risk and expected stock returns: When does the tail wag the dog? *The Review of Asset Pricing Studies*, 4(2):206–246.

Banerjee, A. V. (1992). A simple model of herd behavior. *The quarterly journal of economics*, 107(3):797–817.

Barberis, N. and Thaler, R. (2003). A survey of behavioral finance. *Handbook of the Economics of Finance*, 1:1053–1128.

Barro, R. J. (2006). Rare disasters and asset markets in the twentieth century. *The Quarterly Journal of Economics*, 121(3):823–866.

Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.

Blasco, N., Corredor, P., and Ferreruela, S. (2012). Does herding affect volatility? implications for the spanish stock market. *Quantitative Finance*, 12(2):311–327.

Bollerslev, T. and Todorov, V. (2011). Tails, fears, and risk premia. *The Journal of Finance*, 66(6):2165–2211.

Bouchaud, J.-P. (2008). Economics needs a scientific revolution. *Nature*, 455(7217):1181–1181.

Bouri, E., Gupta, R., and Roubaud, D. (2019). Herding behaviour in cryptocurrencies. *Finance Research Letters*, 29:216–221.

Caillault, C. and Guégan, D. (2005). Empirical estimation of tail dependence using copulas: application to Asian markets. *Quantitative finance*, 5(5):489–501.

Campbell, J. Y. (1996). Understanding risk and return. *Journal of Political economy*, 104(2):298–345.

Capéraà, P., Fougères, A.-L., and Genest, C. (1997). A nonparametric estimation procedure for bivariate extreme value copulas. *Biometrika*, 84(3):567–577.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance*, 52(1):57–82.

Castro-Camilo, D. and Huser, R. (2020). Local likelihood estimation of complex tail dependence structures, applied to us precipitation extremes. *Journal of the American Statistical Association*, 115(531):1037–1054.

Chabi-Yo, F., Huggenberger, M., and Weigert, F. (2021). Multivariate crash risk. *Journal of Financial Economics*.

Chabi-Yo, F., Ruenzi, S., and Weigert, F. (2018). Crash sensitivity and the cross section of expected stock returns. *Journal of Financial and Quantitative Analysis*, 53(3):1059–1100.

Chang, E. C., Cheng, J. W., and Khorana, A. (2000). An examination of herd behavior in equity markets: An international perspective. *Journal of Banking & Finance*, 24(10):1651–1679.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American statistical association*, 91(434):862–872.

Chen, C. Y.-H., Härdle, W. K., and Klochkov, Y. (2021). Sonic: Social network analysis with influencers and communities. *Journal of Econometrics*.

Chen, H., De, P., Hu, Y. J., and Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, 27(5):1367–1403.

Cipriani, M. and Guarino, A. (2014). Estimating a structural model of herd behavior in financial markets. *American Economic Review*, 104(1):224–51.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of finance*, 66(4):1047–1108.

Colander, D., Föllmer, H., Haas, A., Goldberg, M. D., Juselius, K., Kirman, A., Lux, T., and Sloth, B. (2009). The financial crisis and the systemic failure of academic economics. *Univ. of Copenhagen Dept. of Economics Discussion Paper*, (09-03).

Coles, S., Heffernan, J., and Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365.

Cont, R. and Bouchaud, J.-P. (2000). Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic dynamics*, 4(2):170–196.

da Gama Silva, P. V. J., Klotzle, M. C., Pinto, A. C. F., and Gomes, L. L. (2019). Herding behavior and contagion in the cryptocurrency market. *Journal of Behavioral and Experimental Finance*, 22:41–50.

Daniel, K. and Titman, S. (1997). Evidence on the characteristics of cross sectional variation in stock returns. *the Journal of Finance*, 52(1):1–33.

Das, S. R. and Chen, M. Y. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management science*, 53(9):1375–1388.

De Luca, G. and Zuccolotto, P. (2011). A tail dependence-based dissimilarity measure for financial time series clustering. *Advances in data analysis and classification*, 5(4):323–340.

Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. *Bulletins de l'académie royale de Belgique*, 65(1):274–292.

Deng, A. (2014). Understanding spurious regression in financial economics. *Journal of Financial Econometrics*, 12(1):122–150.

Dittmar, R. F. (2002). Nonlinear pricing kernels, kurtosis preference, and evidence from the cross section of equity returns. *The Journal of Finance*, 57(1):369–403.

Einmahl, J., Krajina, A., and Segers, J. (2008). A method of moments estimator of tail dependence. *Bernoulli*, 14(4):1003–1026.

Fagiolo, G., Moneta, A., and Windrum, P. (2007). A critical guide to empirical validation of agent-based models in economics: Methodologies, procedures, and open problems. *Computational Economics*, 30(3):195–226.

Fama, E. and French, K. (1996). Multifactor explanations of asset pricing anomalies. *Journal of Finance*, 51(1):55–84.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417.

Fama, E. F. and French, K. R. (1995). Size and book-to-market factors in earnings and returns. *The journal of finance*, 50(1):131–155.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of political economy*, 81(3):607–636.

Fang, H. and Lai, T.-Y. (1997). Co-kurtosis and capital asset pricing. *Financial Review*, 32(2):293–307.

Fermanian, J.-D., Radulovic, D., and Wegkamp, M. (2004). Weak convergence of empirical copula processes. *Bernoulli*, 10(5):847–860.

Frahm, G., Junker, M., and Schmidt, R. (2005a). Estimating the tail-dependence coefficient: properties and pitfalls. *Insurance: mathematics and Economics*, 37(1):80–100.

Frahm, G., Junker, M., and Schmidt, R. (2005b). Estimating the tail-dependence coefficient: properties and pitfalls. *Insurance: mathematics and economics*, 37(1):80–100.

Franke, R. (2008). A microfounded herding model and its estimation on german survey expectations. *European Journal of Economics and Economic Policies: Intervention*, 5(2):301–328.

Frazzini, A. and Pedersen, L. H. (2014). Betting against beta. *Journal of Financial Economics*, 111(1):1–25.

Froot, K. A., Scharfstein, D. S., and Stein, J. C. (1992). Herd on the street: Informational inefficiencies in a market with short-term speculation. *The Journal of Finance*, 47(4):1461–1484.

Galambos, J. (1978). The asymptotic theory of extreme order statistics. Technical report.

Garcin, M. (2017). Estimation of time-dependent Hurst exponents with variational smoothing and application to forecasting foreign exchange rates. *Physica A: statistical mechanics and its applications*, 483:462–479.

Garcin, M. and Goulet, C. (2019). Non-parametric news impact curve: a variational approach. *Soft computing*, 24:13797–13812.

Garcin, M. and Guégan, D. (2012). Extreme values of random or chaotic discretization steps and connected networks. *Applied mathematical sciences*, 6(119):5901–5926.

Garcin, M. and Guégan, D. (2016). Wavelet shrinkage of a noisy dynamical system with non-linear noise impact. *Physica D: nonlinear phenomena*, 325:126–145.

Garcin, M., Guégan, D., and Hassani, B. (2021). A multivariate quantile based on Kendall ordering. *to appear in Revstat - statistical journal*.

Genest, C. and Favre, A. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering*, 12(4):347–368.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.

Genest, C. and Rémillard, B. (2004). Test of independence and randomness based on the empirical copula process. *Test*, 13(2):335–369.

Genest, C. and Rivest, L.-P. (1989). A characterization of Gumbel's family of extreme value distributions. *Statistics & probability letters*, 8(3):207–211.

Ghonghadze, J. and Lux, T. (2011). Modelling the dynamics of eu economic sentiment indicators: An interaction-based approach. *Applied Economics*, 44(24):3065–3088.

Guo, K., Sun, Y., and Qian, X. (2017). Can investor sentiment be used to predict the stock price? dynamic analysis based on china stock market. *Physica A: Statistical Mechanics and its Applications*, 469:390–396.

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2012). *Nonparametric and semiparametric models*. Springer science and business media.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). . . . and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68.

Harvey, C. R. and Siddique, A. (2000). Conditional skewness in asset pricing tests. *The Journal of finance*, 55(3):1263–1295.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The annals of statistics*, pages 1163–1174.

Hoga, Y. (2018). A structural break test for extremal dependence in β-mixing random vectors. *Biometrika*, 105(3):627–643.

Huang, W., Liu, Q., Rhee, S. G., and Wu, F. (2012). Extreme downside risk and expected stock returns. *Journal of Banking & Finance*, 36(5):1492–1502.

Hurn, A. S., Jeisman, J., and Lindsay, K. A. (2007). Seeing the wood for the trees: A critical evaluation of methods to estimate the parameters of stochastic differential equations. *Journal of Financial Econometrics*, 5(3):390–455.

Huser, R. and Wadsworth, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *Journal of the American Statistical Association*, 114(525):434–444.

Idier, J., Lamé, G., and Mésonnier, J.-S. (2014). How useful is the marginal expected shortfall for the measurement of systemic exposure? a practical assessment. *Journal of Banking & Finance*, 47:134–146.

Iori, G. (2002). A microsimulation of traders activity in the stock market: the role of heterogeneity, agents' interactions and trade frictions. *Journal of Economic Behavior & Organization*, 49(2):269–285.

Jensen, B. and Poulsen, R. (2002). Transition densities of diffusion processes: numerical comparison of approximation techniques. *The Journal of Derivatives*, 9(4):18–32.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC press.

Joe, H. (2014). *Dependence modeling with copulas*. CRC press.

Joe, H., Smith, R., and Weissman, I. (1992). Bivariate threshold methods for extremes. *Journal of the royal statistical society: series B (methodological)*, 54(1):171–183.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American statistical association*, 91(433):401–407.

Juri, A. and Wüthrich, M. (2002). Copula convergence theorems for tail events. *Insurance: mathematics and economics*, 30(3):405–420.

Juri, A. and Wüthrich, M. (2003). Tail dependence from a distributional point of view. *Extremes*, 6(3):213–246.

Kelly, B. and Jiang, H. (2014). Tail risk and asset prices. *The Review of Financial Studies*, 27(10):2841–2871.

Keynes, J. M. (1936). *The general theory of employment, interest, and money*. Springer.

Keynes, J. M. (1937). The general theory of employment. *The quarterly journal of economics*, 51(2):209–223.

Kim, S.-H. and Kim, D. (2014). Investor sentiment from internet message postings and the predictability of stock returns. *Journal of Economic Behavior & Organization*, 107:708–729.

Kirman, A. (1991). Epidemics of opinion and speculative bubbles in financial matkets. *Money and financial markets chap*, 17.

Kirman, A. (1993). Ants, rationality, and recruitment. *The Quarterly Journal of Economics*, 108(1):137–156.

Klüppelberg, C., Kuhn, G., and Peng, L. (2007). Estimating the tail dependence function of an elliptical distribution. *Bernoulli*, 13(1):229–251.

Koltchinskii, V. (1997). M-estimation, convexity and quantiles. *Annals of statistics*, 25(2):435–477.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Lewellen, J., Nagel, S., and Shanken, J. (2010). A skeptical appraisal of asset pricing tests. *Journal of Financial economics*, 96(2):175–194.

Longin, F. and Solnik, B. (2001). Extreme correlation of international equity markets. *Journal of finance*, 56(2):649–676.

Lu, Z. and Murray, S. (2019). Bear beta. *Journal of Financial Economics*, 131(3):736–760.

Lux, T. (1995). Herd behaviour, bubbles and crashes. *The economic journal*, 105(431):881–896.

Lux, T. (1998). The socio-economic dynamics of speculative markets: interacting agents, chaos, and the fat tails of return distributions. *Journal of Economic Behavior & Organization*, 33(2):143–165.

Lux, T. (2009). Rational forecasts or social opinion dynamics? identification of interaction effects in a business climate survey. *Journal of Economic Behavior & Organization*, 72(2):638–655.

Lux, T. (2012). Estimation of an agent-based model of investor sentiment formation in financial markets. *Journal of Economic Dynamics and Control*, 36(8):1284–1302.

Lux, T. (2018). Estimation of agent-based models using sequential monte carlo methods. *Journal of Economic Dynamics and Control*, 91:391–408.

Lux, T. and Marchesi, M. (1999). Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature*, 397(6719):498–500.

Lux, T. and Marchesi, M. (2000). Volatility clustering in financial markets: a microsimulation of interacting agents. *International journal of theoretical and applied finance*, 3(04):675–702.

Lux, T. and Zwinkels, R. C. (2018). Empirical validation of agent-based models. In *Handbook of*

computational economics, volume 4, pages 437–488. Elsevier.

Mahmoudi, N., Docherty, P., and Moscato, P. (2018). Deep neural networks understand investors better. *Decision Support Systems*, 112:23–34.

Malevergne, Y. and Sornette, D. (2003). Testing the Gaussian copula hypothesis for financial assets dependences. *Quantitative finance*, 3:231–250.

Mallat, S. (1999). *A wavelet tour of signal processing*. Elsevier.

Mandelbrot, B. (1963). The variation of certain speculative prices. *The Journal of Business*, 36(4):394–419.

McLean, R. D. and Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, 71(1):5–32.

Meine, C., Supper, H., and Weiß, G. N. (2016). Is tail risk priced in credit default swap premia? *Review of Finance*, 20(1):287–336.

Nelder, J. and Mead, R. (1965). A simplex method for function minimization. *Computer journal*, 7(4):308–313.

Nelsen, R. (2007). *An introduction to copulas*. Springer science & business media.

Neuhäuser, D., Hirsch, C., Gloaguen, C., and Schmidt, V. (2015). Joint distributions for total lengths of shortest-path trees in telecommunication networks. *Annals of telecommunications*, 70(5-6):221–232.

Newey, W. K., West, K. D., et al. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Nicolas, M. L. D. (2022). pycop: a Python package for dependence modeling with copulas.

Oliveira, N., Cortez, P., and Areal, N. (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, 85:62–73.

Orléan, A. (1995). Bayesian interactions and collective dynamics of opinion: Herd behavior and mimetic contagion. *Journal of Economic Behavior & Organization*, 28(2):257–274.

Patton, A. (2004). On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. *Journal of financial econometrics*, 2(1):130–168.

Poon, S.-H., Rockinger, M., and Tawn, J. (2004). Extreme value dependence in financial markets: diagnostics, models, and financial implications. *Review of financial studies*, 17(2):581–610.

Poulin, A., Huard, D., Favre, A.-C., and Pugin, S. (2007). Importance of tail dependence in bivariate frequency analysis. *Journal of hydrologic engineering*, 12(4):394–403.

Ranta, M. (2010). *Wavelet multiresolution analysis of financial time series*. PhD thesis, Universitas Wasaensis.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the us stock market. *Journal of Banking & Finance*, 84:25–40.

Renault, T. (2020). Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. *Digital Finance*, 2(1):1–13.

Rietz, T. A. (1988). The equity risk premium a solution. *Journal of monetary Economics*, 22(1):117–131.

Roy, A. D. (1952). Safety first and the holding of assets. *Econometrica: Journal of the econometric society*, pages 431–449.

Ruenzi, S., Ungeheuer, M., and Weigert, F. (2020). Joint extreme events in equity returns and liquidity and their cross-sectional pricing implications. *Journal of Banking & Finance*, 115:105809.

Ruenzi, S. and Weigert, F. (2018). Momentum and crash sensitivity. *Economics Letters*, 165:77–81.

Sato, M., Ichiki, K., and Takeuchi, T. (2011). Copula cosmology: Constructing a likelihood function. *Physical review D*, 83(2):203501.

Scharfstein, D. S. and Stein, J. C. (1990). Herd behavior and investment. *The American economic review*, pages 465–479.

Schepsmeier, U. and Stöber, J. (2014). Web supplement: Derivatives and Fisher information of bivariate copulas. *Statistical papers*, 55(2):525–542.

Scherrer, R., Berlind, A., Mao, Q., and McBride, C. (2009). From finance to cosmology: The copula of large-scale structure. *Astrophysical journal letters*, 708(1):L9.

Schmidt, R. and Stadtmüller, U. (2006). Non-parametric estimation of tail dependence. *Scandinavian journal of statistics*, 33(2):307–335.

Schmidt, R. and Stadtmüller, U. (2006). Non-parametric estimation of tail dependence. *Scandinavian journal of statistics*, 33(2):307–335.

Serinaldi, F. (2008). Analysis of inter-gauge dependence by Kendall's τ_k, upper tail dependence coefficient, and 2-copulas with application to rainfall fields. *Stochastic environmental research and risk assessment*, 22(6):671–688.

Shahzad, S. J. H., Anas, M., and Bouri, E. (2022). Price explosiveness in cryptocurrencies and elon musk's tweets. *Finance Research Letters*, page 102695.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442.

Shi, Y., Tang, Y.-r., and Long, W. (2019). Sentiment contagion analysis of interacting investors: evidence from china's stock forum. *Physica A: Statistical Mechanics and its Applications*, 523:246–259.

Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51:1384–1399.

Shiller, R. J. (2015). Irrational exuberance. In *Irrational exuberance*. Princeton university press.

Shiller, R. J., Fischer, S., and Friedman, B. M. (1984). Stock prices and social dynamics. *Brookings papers on economic activity*, 1984(2):457–510.

Sibuya, M. (1960). Bivariate extreme statistics, I. *Annals of the institute of statistical mathematics*, 11(3):195–210.

Sklar, M. (1959). Fonctions de répartition à *n* dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231.

Smith, R. L., Tawn, J. A., and Coles, S. G. (1997). Markov chain models for threshold exceedances. *Biometrika*, 84(2):249–268.

Sornette, D. (2017). Why stock markets crash. In *Why Stock Markets Crash*. Princeton university press.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., and Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20(5):926–957.

Supper, H., Irresberger, F., and Weiß, G. (2020). A comparison of tail dependence estimators. *European journal of operational research*, 284(2):728–742.

Tawn, J. A. (1988). Bivariate extreme value theory: models and estimation. *Biometrika*, 75(3):397–415.

Topol, R. (1991). Bubbles and volatility of stock prices: effect of mimetic contagion. *The Economic Journal*, 101(407):786–800.

Van Oordt, M. R. and Zhou, C. (2016). Systematic tail risk. *Journal of Financial and Quantitative Analysis*, 51(2):685–705.

Wang, G. and Wang, Y. (2018). Herding, social network and volatility. *Economic Modelling*, 68:74–81.

Weidlich, W. (1971). The statistical description of polarization phenomena in society. *British Journal of Mathematical and Statistical Psychology*, 24(2):251–266.

Weigert, F. (2016). Crash aversion and the cross-section of expected stock returns worldwide. *The Review of Asset Pricing Studies*, 6(1):135–178.

Zheng, B., Ren, F., Trimper, S., and Zheng, D. (2004). A generalized dynamic herding model with feed-back interactions. *Physica A: Statistical Mechanics and its Applications*, 343:653–661.